Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis

Wenmeng Yu, Hua Xu, * Ziqi Yuan, Jiele Wu

State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
ywm18@mails.tsinghua.edu.cn, xuhua@mail.tsinghua.edu.cn,
ziqiyuan@bupt.edu.cn, 1120171196@bit.edu.cn

Abstract

Representation Learning is a significant and challenging task in multimodal learning. Effective modality representations should contain two parts of characteristics: the consistency and the difference. Due to the unified multimodal annotation, existing methods are restricted in capturing differentiated information. However, additional unimodal annotations are high time- and labor-cost. In this paper, we design a label generation module based on the self-supervised learning strategy to acquire independent unimodal supervisions. Then, joint training the multimodal and uni-modal tasks to learn the consistency and difference, respectively. Moreover, during the training stage, we design a weight-adjustment strategy to balance the learning progress among different subtasks. That is to guide the subtasks to focus on samples with the larger difference between modality supervisions. Last, we conduct extensive experiments on three public multimodal baseline datasets. The experimental results validate the reliability and stability of auto-generated unimodal supervisions. On MOSI and MOSEI datasets, our method surpasses the current state-of-the-art methods. On the SIMS dataset, our method achieves comparable performance than human-annotated unimodal labels. The full codes are available at https://github.com/thuiar/Self-MM.

Introduction

Multimodal Sentiment Analysis (MSA) attracts more and more attention in recent years (Zadeh et al. 2017; Tsai et al. 2019; Poria et al. 2020). Compared with unimodal sentiment analysis, multimodal models are more robust and achieve salient improvements when dealing with social media data. With the booming of user-generated online content, MSA has been introduced into many applications such as risk management, video understanding, and video transcription.

Though previous works have made impressive improvements on benchmark datasets, MSA is still full of challenges. Baltrušaitis, Ahuja, and Morency (2019) identified five core challenges for multimodal learning: alignment, translation, representation, fusion, and co-learning. Among them, representation learning stands in a fundamental position. In most recent work, Hazarika, Zimmermann, and Poria (2020) stated that unimodal representations should contain both consistent and complementary information. According to the difference of guidance in representation learning, we classify existing methods into the two categories: forward guidance and backward guidance. In forward-guidance methods, researches are devoted to design interactive modules for capturing the cross-modal information (Zadeh et al. 2018a; Sun et al. 2020; Tsai et al. 2019). However, due to the unified multimodal annotation, it is difficult for them to capture modality-specific information. Shown in Figure 1, the unified multimodal labels are not always suitable for the unimodal learning. In backward-guidance methods, researches proposed additional loss function as prior constraint, which leads modality representations to contain both consistent and complementary information (Yu et al. 2020a; Hazarika, Zimmermann, and Poria 2020). However, the former needed additional labor costs, and the latter were difficult to represent the modality-specific difference with spatial differences.

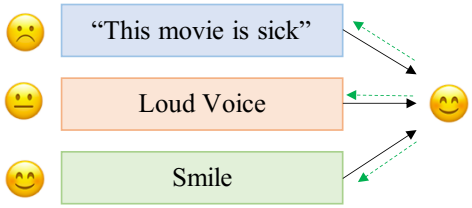

Figure 1: An example of unimodal labels and multimodal labels, from Zadeh et al. (2017). The green dotted lines represent the process of backpropagation.

In this paper, we focus on the backward-guidance method. Motivated by the independent unimodal annotations and advanced modality-specific representation learning, we propose a novel self-supervised multi-task learning strategy. Different from Yu et al. (2020a), our method does not need human-annotated unimodal labels but uses auto-generated unimodal labels. It is based on two intuitions. First, label difference is positively correlated with the distance difference between modality representations and class centers. Second, unimodal labels are highly related to multimodal labels. Hence, we design a unimodal label generation module based on multimodal labels and modality representations. The details are shown in Section .

*Hua Xu is the corresponding author

Considering that auto-generated unimodal labels are not stable enough at the beginning epochs, we design a momentum-based update method, which applies a larger weight for the unimodal labels generated later. Furthermore, we introduce a self-adjustment strategy to adjust each subtask's weight when integrating the final multi-task loss function. We believe that it is difficult for subtasks with small label differences, between auto-generated unimodal labels and human-annotated multimodal labels, to learn the modality-specific representations. Therefore, the weight of subtasks is positively correlated with the labels difference.

The novel contributions of our work can be summarized as follows:

- We propose the relative distance value based on the distance between modality representations and class centers, positively correlated with model outputs.
- We design a unimodal label generation module based on the self-supervised strategy. Furthermore, a novel weight self-adjusting strategy is introduced to balance different task loss constraints.
- Extensive experiments on three benchmark datasets validate the stability and reliability of auto-generated unimodal labels. Moreover, our method outperforms current state-of-the-art results.

Related Work

In this section, we mainly discuss related works in the domain of multimodal sentiment analysis and multi-task learning. We also emphasize the innovation of our work.

Multimodal Sentiment Analysis

Multimodal sentiment analysis has become a significant research topic that integrates verbal and nonverbal information like visual and acoustic. Previous researchers mainly focus on representation learning and multimodal fusion. For representation learning methods, Wang et al. (2019) constructed a recurrent attended variation embedding network to generate multimodal shifting. Hazarika, Zimmermann, and Poria (2020) designed two distinct encoders projecting each modality into modality-invariant and modality-specific space. Two regularization components are claimed to aid modality-invariant and modality-specific representation learning. Yu et al. (2020a) introduced independent unimodal human annotations. By joint learning unimodal and multimodal tasks, the proposed multi-task multimodal framework learned modality-specific and modality-invariant representations simultaneously. For multimodal fusion, according to the fusion stage, previous works can be classified into two categories: early fusion and late fusion. Early fusion methods usually use delicate attention mechanisms for cross-modal fusion. Zadeh et al. (2018a) designed a memory fusion network for cross-view interactions. Tsai et al. (2019) proposed cross-modal transformers, which learn the cross-modal attention to reinforce a target modality. Late fusion methods learn intra-modal representation first and perform inter-modal fusion last. Zadeh et al. (2017) used a tensor fusion network that obtains tensor representation by computing the outer product between unimodal representations. Liu et al. (2018) proposed a low-rank multimodal fusion method to decrease the computational complexity of tensor-based methods.

Our work aims at representation learning based on late fusion structure. Different from previous studies, we joint learn unimodal and multimodal tasks with the self-supervised strategy. Our method learns similarity information from multimodal task and learns differentiated information from unimodal tasks.

Transformer and BERT

Transformer is a sequence-to-sequence architecture without recurrent structure (Vaswani et al. 2017). It is used for modeling sequential data and has superior performance on results, speed, and depth than recurrent structure. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018) is a successful application on the transformer. The pre-trained BERT model has achieved significant improvements in multiple NLP tasks. In multimodal learning, pre-trained BERT also achieved remarkable results. Currently, there are two ways to use pre-trained BERT. The first is to use the pre-trained BERT as a language feature extraction module (Hazarika, Zimmermann, and Poria 2020). The second is to integrate acoustic and visual information on the middle layers (Tsai et al. 2019; Rahman et al. 2020). In this paper, we use the first way and finetune the pre-trained BERT for our tasks.

Multi-task Learning

Multi-task learning aims to improve the generalization performance of multiple related tasks by utilizing the knowledge contained in different tasks (Zhang and Yang 2017). Compared with single-task learning, there are two main challenges for multi-task learning in the training stage. The first is how to share network parameters, including hard-sharing and soft-sharing methods. The second is how to balance the learning process of different tasks. Recently, multi-task learning is wildly applied in MSA (Liu et al. 2015; Zhang et al. 2016; Akhtar et al. 2019; Yu et al. 2020b).

In this work, we introduce unimodal subtasks to aid the modality-specific representation learning. We adopt a hard-sharing strategy and design a weight-adjustment method to solve the problem of how to balance.

Methodology

In this section, we explain the Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) in detail. The goal of the Self-MM is to acquire information-rich unimodal representations by joint learning one multimodal task and three unimodal subtasks. Different from the multimodal task, the labels of unimodal subtasks are auto-generated in the self-supervised method. For the convenience of the following sections, we refer the human-annotated multimodal labels as **m-labels** and the auto-generated unimodal labels as **u-labels**.

Task Setup

Multimodal Sentiment Analysis (MSA) is to judge the sentiments using multimodal signals, including text (I_t), audio

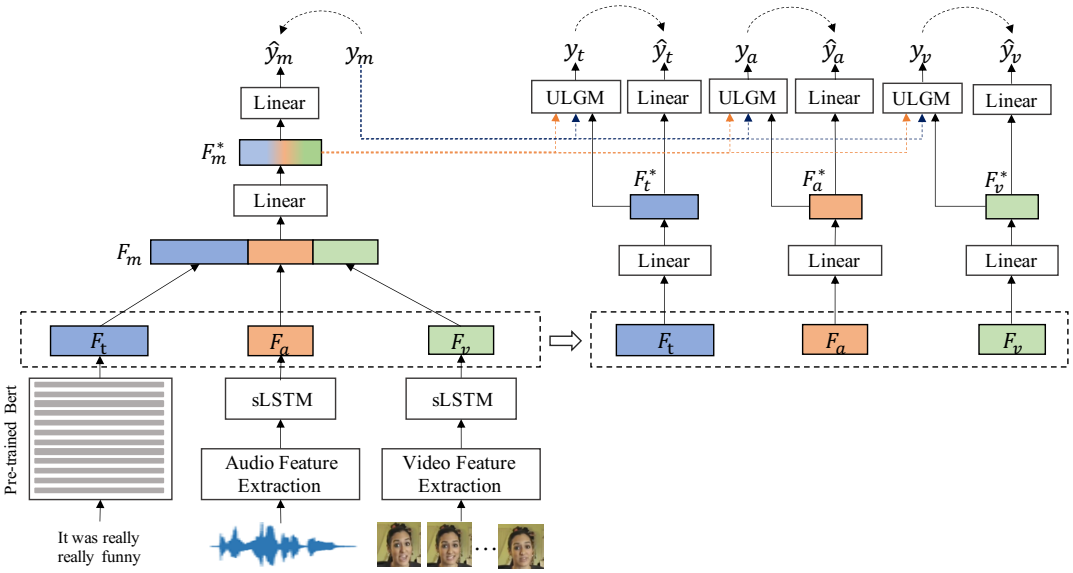

Figure 2: The overall architecture of Self-MM. The $\hat{y}_m$, $\hat{y}_t$, $\hat{y}_a$, and $\hat{y}_v$ are the predictive outputs of the multimodal task and the three unimodal tasks, respectively. The y_m is the multimodal annotation by human. The y_t, y_a, and y_v are the unimodal supervision generated by the self-supervised strategy. Finally, $\hat{y}_m$ is used as the sentiment output.

(I_a), and vision (I_v). Generally, MSA can be regarded as either a regression task or a classification task. In this work, we regard it as the regression task. Therefore, Self-MM takes $I_t, I_a,$ andI_v as inputs and outputs one sentimental intensity result $\hat{y}_m \in R$. In the training stage, to aid representation learning, Self-MM has extra three unimodal outputs $\hat{y}_s \in R$, where $s \in \{t, a, v\}$. Though more than one output, we only use $\hat{y}_m$ as the final predictive result.

Overall Architecture

Shown in Figure 2, the Self-MM consists of one multimodal task and three independent unimodal subtasks. Between the multimodal task and different unimodal tasks, we adopt hard-sharing strategy to share the bottom representation learning network.

Multimodal Task. For the multimodal task, we adopt a classical multimodal sentiment analysis architecture. It contains three main parts: the feature representation module, the feature fusion module, and the output module. In the text modality, since the great success of the pre-trained language model, we use the pre-trained 12-layers BERT to extract sentence representations. Empirically, the first-word vector in the last layer is selected as the whole sentence representation F_t.

$$F_t = BERT(I_t; \theta_t^{bert}) \in R^{d_t}$$

In audio and vision modalities, following Zadeh et al. (2017); Yu et al. (2020b), we use pre-trained ToolKits to extract the initial vector features, $I_a \in R^{l_a \times d_a}$ and $I_v \in R^{l_v \times d_v}$, from raw data. Here, l_a and l_v are the sequence lengths of audio and vision, respectively. Then, we use a single directional Long Short-Term Memory (sLSTM) (Hochreiter and Schmidhuber 1997) to capture the timing characteristics. Finally, the end-state hidden vectors are adopted as the whole sequence representations.

$$F_a = sLSTM(I_a; \theta_a^{lstm}) \in R^{d_a}$$
$$F_v = sLSTM(I_v; \theta_v^{lstm}) \in R^{d_v}$$

Then, we concatenate all uni-modal representations and project them into a lower-dimensional space R^{d_m}.

$$F_m^* = ReLU(W_{l1}^{m\,T}[F_t; F_a; F_v] + b_{l1}^m)$$

where $W_{l1}^m \in R^{(d_t+d_a+d_v)\times d_m}$ and $ReLU$ is the relu activation function.

Last, the fusion representation F_m^* is used to predict the multimodal sentiment.

$$\hat{y}_m = W_{l2}^{m\,T} F_m^* + b_{l2}^m$$

where $W_{l2}^m \in R^{d_m \times 1}$.

Uni-modal Task. For the three unimodal tasks, they share modality representations with the multimodal task. In order to reduce the dimensional difference between different modalities, we project them into a new feature space. Then, get the uni-modal results with linear regression.

$$F_s^* = ReLU(W_{l1}^{s\,T} F_s + b_{l1}^s)$$
$$\hat{y}_s = W_{l2}^{s\,T} F_s^* + b_{l2}^s$$

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

To guide the unimodal tasks' training process, we design a Unimodal Label Generation Module (ULGM) to get u-labels. Details of the ULGM are discussed in Section .

$$y_s = ULGM(y_m, F_m^*, F_s^*)$$

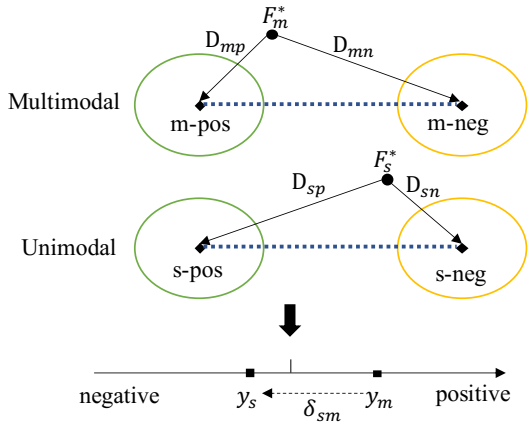

Figure 3: Unimodal label generation example. Multimodal representation F_m^* is closer to the positive center (m-pos) while unimodal representation is closer to the negative center (s-neg). Therefore, unimodal supervision y_s is added a negative offset δ_{sm} to the multimodal label y_m

.

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

Finally, we joint learn the multimodal task and three unimodal tasks under m-labels and u-labels supervision. It is worth noting that these unimodal tasks are only exist in the training stage. Therefore, we use $\hat{y}_m$ as the final output.

ULGM

The ULGM aims to generate unimodal supervision values based on multimodal annotations and modality representations. In order to avoid unnecessary interference with the update of network parameters, the ULGM is designed as a non-parameter module. Generally, unimodal supervision values are highly correlated with multimodal labels. Therefore, the ULGM calculates the offset according to the relative distance from modality representations to class centers, shown as Figure 3.

Relative Distance Value. Since different modality representations exist in different feature spaces, using the absolute distance value is not accurate enough. Therefore, we propose the relative distance value, which is not related to the space difference. First, when in training process, we maintain the positive center (C_i^p) and the negative center (C_i^n) of different modality representations:

$$C_i^p = \frac{\sum_{j=1}^{N} I(y_i(j)>0) \cdot F_{ij}^g}{\sum_{j=1}^{N} I(y_i(j)>0)} \quad (1)$$

$$C_i^n = \frac{\sum_{j=1}^{N} I(y_i(j)<0) \cdot F_{ij}^g}{\sum_{j=1}^{N} I(y_i(j)<0)} \quad (2)$$

where $i \in \{m, t, a, v\}$, N is the number of training samples, and $I(\cdot)$ is a indicator function. F_{ij}^g is the global representation of the j_{th} sample in modality i.

For modality representations, we use L2 normalization as the distance between F_i^* and class centers.

$$D_i^p = \frac{||F_i^* - C_i^p||_2^2}{\sqrt{d_i}} \quad (3)$$

$$D_i^n = \frac{||F_i^* - C_i^n||_2^2}{\sqrt{d_i}} \quad (4)$$

Algorithm 1 Unimodal Supervisions Update Policy

Input: unimodal inputs I_t, I_a, I_v, m-labels y_m
Output: u-labels $y_t^{(i)}, y_a^{(i)}, y_v^{(i)}$ where i means the number of training epochs

```
Initialize model parameters M(θ; x)
Initialize u-labels y_t^(1) = y_m, y_a^(1) = y_m, y_v^(1) = y_m
Initialize global representations F_t^g = 0, F_a^g = 0, F_v^g = 0, F_m^g = 0
for n ∈ [1, end] do
    for mini-batch in dataLoader do
        Compute mini-batch modality representations F_t^*, F_a^*, F_v^*, F_m^*
        Compute loss L using Equation (10)
        Compute parameters gradient ∂L/∂θ
        Update model parameters: θ = θ − η ∂L/∂θ
        if n ≠ 1 then
            Compute relative distance values α_m, α_t, α_a, and α_v using Equation (1∼5)
            Compute y_t, y_a, y_v using Equation (8)
            Update y_t^(n), y_a^(n), y_t^(n) using Equation (9)
        end if
        Update global representations F_s^g using F_s^*, where s ∈ {m, t, a, v}
    end for
end for
```

where $i \in \{m, t, a, v\}$. d_i is the representation dimension, a scale factor.

Then, we define the relative distance value, which evaluates the relative distance from the modality representation to the positive center and the negative center.

$$\alpha_i = \frac{D_i^n - D_i^p}{D_i^p + \epsilon} \quad (5)$$

where $i \in \{m, t, a, v\}$. ϵ is a small number in case of zero exceptions.

Shifting Value. It is intuitive that α_i is positively related to the final results. To get the link between supervisions and predicted values, we consider the following two relationships.

$$\frac{y_s}{y_m} \propto \frac{\hat{y}_s}{\hat{y}_m} \propto \frac{\alpha_s}{\alpha_m} \Rightarrow y_s = \frac{\alpha_s * y_m}{\alpha_m} \quad (6)$$

$$y_s - y_m \propto \hat{y}_s - \hat{y}_m \propto \alpha_s - \alpha_m \Rightarrow y_s = y_m + \alpha_s - \alpha_m \quad (7)$$

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

Specifically, the Equation 7 is introduced to avoid the "zero value problem". In Equation 6, when y_m equals to zero, the generated unimodal supervision values y_s are always zero. Then, joint considering the above relationships, we can get unimodal supervisions by equal-weight summation.

$$\begin{aligned} y_s &= \frac{y_m * \alpha_s}{2\alpha_m} + \frac{y_m + \alpha_s - \alpha_m}{2} \\ &= y_m + \frac{\alpha_s - \alpha_m}{2} * \frac{y_m + \alpha_m}{\alpha_m} \\ &= y_m + \delta_{sm} \end{aligned} \quad (8)$$

Model	MOSI				MOSEI				Data Setting
	MAE	Corr	Acc-2	F1-Score	MAE	Corr	Acc-2	F1-Score	
TFN (B)[1]	0.901	0.698	-/80.8	-/80.7	0.593	0.700	-/82.5	-/82.1	Unaligned
LMF (B)[1]	0.917	0.695	-/82.5	-/82.4	0.623	0.677	-/82.0	-/82.1	Unaligned
MFN[1]	0.965	0.632	77.4/-	77.3/-	-	-	76.0/-	76.0/-	Aligned
RAVEN[1]	0.915	0.691	78.0/-	76.6/-	0.614	0.662	79.1/-	79.5/-	Aligned
MFM (B)[1]	0.877	0.706	-/81.7	-/81.6	0.568	0.717	-/84.4	-/84.3	Aligned
MulT (B)[1]	0.861	0.711	81.5/84.1	80.6/83.9	0.58	0.703	-/82.5	-/82.3	Aligned
MISA (B)[1]	0.783	0.761	81.8/83.4	81.7/83.6	0.555	0.756	83.6/85.5	83.8/85.3	Aligned
MAG-BERT (B)[2]	0.712	0.796	84.2/86.1	84.1/86.0	-	-	84.7/-	84.5/-	Aligned
MISA (B)*	0.804	0.764	80.79/82.1	80.77/82.03	0.568	0.724	82.59/84.23	82.67/83.97	Aligned
MAG-BERT (B)*	0.731	0.789	82.54/84.3	82.59/84.3	0.539	0.753	**83.79/85.23**	**83.74**/85.08	Aligned
Self-MM (B)*	**0.713**	**0.798**	**84.00/85.98**	**84.42/85.95**	**0.530**	**0.765**	82.81/85.17	82.53/**85.30**	Unaligned

Table 1: Results on MOSI and MOSEI. (B) means the language features are based on BERT; [1] is from Hazarika, Zimmermann, and Poria (2020) and [2] is from Rahman et al. (2020). Models with * are reproduced under the same conditions. In Acc-2 and F1-Score, the left of the "/" is calculated as "negative/non-negative" and the right is calculated as "negative/positive".

where $s \in \{t, a, v\}$. The $\delta_{sm} = \frac{\alpha_t - \alpha_m}{2} * \frac{y_m + \alpha_m}{\alpha_m}$ represents the offset value of unimodal supervisions to multimodal annotations.

Momentum-based Update Policy. Due to the dynamic changes of modality representations, the generated u-labels calculated by Equation (8) are unstable enough. In order to mitigate the adverse effects, we design a momentum-based update policy, which combines the new generated value with history values.

$$y_s^{(i)} = \begin{cases} y_m & i = 1 \\ \frac{i-1}{i+1} y_s^{(i-1)} + \frac{2}{i+1} y_s^i & i > 1 \end{cases} \quad (9)$$

where $s \in \{t, a, v\}$. y_s^i is the new generated u-labels at the i_{th} epoch. $y_s^{(i)}$ is the final u-labels after the i_{th} epoch.

Formally, assume the total epochs is n, we can get that the weight of y_s^i is $\frac{2i}{(n)(n+1)}$. It means that the weight of u-labels generated later is greater than the previous one. It is in accordance with our experience. Because generated unimodal labels are the cumulative sum of all previous epochs, they will stabilize after enough iterations (about 20 in our experiments). Then, the training process of unimodal tasks will gradually become stable. The unimodal labels update policy is shown in Algorithm 1.

Optimization Objectives

Finally, we use the L1Loss as the basic optimization objective. For uni-modal tasks, we use the difference between u-labels and m-labels as the weight of loss function. It indicates that the network should pay more attention on the samples with larger difference.

$$L = \frac{1}{N} \sum_{i}^{N} (|\hat{y}_m^i - y_m^i| + \sum_{s}^{\{t,a,v\}} W_s^i * |\hat{y}_s^i - y_s^{(i)}|) \quad (10)$$

where N is the number of training samples. $W_s^i = tanh(|y_s^{(i)} - y_m|)$ is the weight of i_{th} sample for auxiliary task s.

Dataset	# Train	# Valid	# Test	# All
MOSI	1284	229	686	2199
MOSEI	16326	1871	4659	22856
SIMS	1368	456	457	2281

Table 2: Dataset statistics in MOSI, MOSEI, and SIMS.

Experimental Settings

In this section, we introduce our experimental settings, including experimental datasets, baselines, and evaluations.

Datasets

In this work, we use three public multimodal sentiment analysis datasets, MOSI (Zadeh et al. 2016), MOSEI (Zadeh et al. 2018b), and SIMS (Yu et al. 2020a). The basic statistics are shown in Table 2. Here, we give a brief introduction to the above datasets.

MOSI. The CMU-MOSI dataset (Zadeh et al. 2016) is one of the most popular benchmark datasets for MSA. It comprises 2199 short monologue video clips taken from 93 Youtube movie review videos. Human annotators label each sample with a sentiment score from -3 (strongly negative) to 3 (strongly positive).

MOSEI. The CMU-MOSEI dataset (Zadeh et al. 2018b) expands its data with a higher number of utterances, greater variety in samples, speakers, and topics over CMU-MOSI. The dataset contains 23,453 annotated video segments (utterances), from 5,000 videos, 1,000 distinct speakers and 250 different topics.

SIMS. The SIMS dataset (Yu et al. 2020a) is a distinctive Chinese MSA benchmark with fine-grained annotations of modality. The dataset consists of 2,281 refined video clips collected from different movies, TV serials, and variety shows with spontaneous expressions, various head poses, occlusions, and illuminations. Human annotators label each sample with a sentiment score from -1 (strongly negative) to

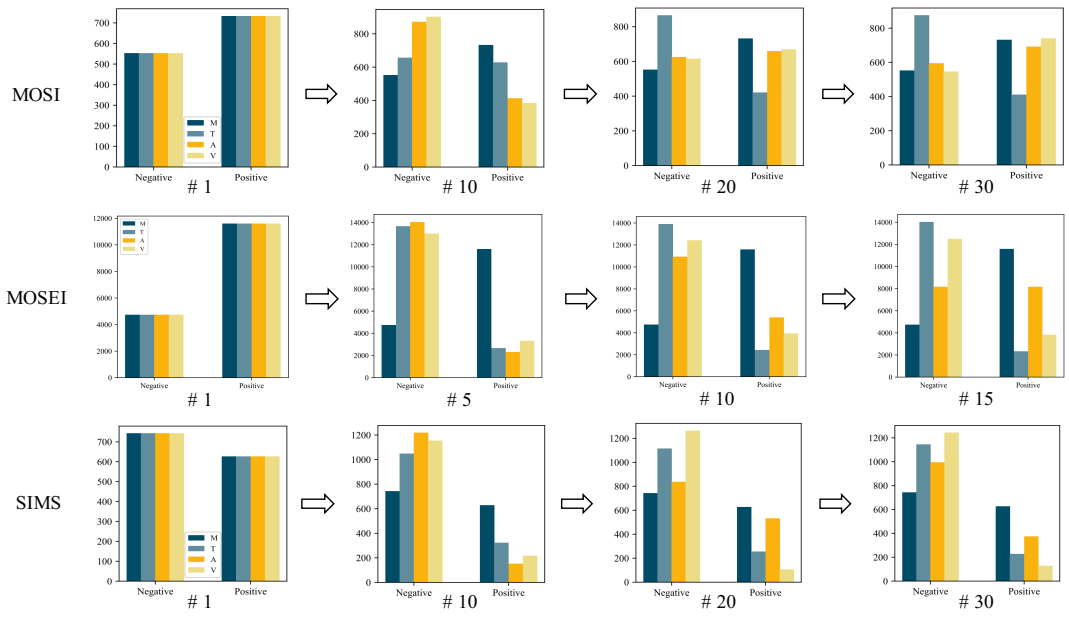

Figure 4: The distribution update process of u-labels on different datasets. The number (#) under each sub picture indicates the number of epochs.

1 (strongly positive).

Baselines

To fully validate the performance of the Self-MM, we make a fair comparison with the following baselines and state-of-the-art models in multimodal sentiment analysis.

TFN. The Tensor Fusion Network (TFN) (Zadeh et al. 2017) calculates a multi-dimensional tensor (based on outer-product) to capture uni-, bi-, and tri-modal interactions.

LMF. The Low-rank Multimodal Fusion (LMF) (Liu et al. 2018) is an improvement over TFN, where low-rank multimodal tensors fusion technique is performed to improve efficiency.

MFN. The Memory Fusion Network (MFN) (Zadeh et al. 2018a) accounts for continuously modeling the view-specific and cross-view interactions and summarizing them through time with a Multi-view Gated Memory.

MFM. The Multimodal Factorization Model (MFM) (Tsai et al. 2018) learns generative representations to learn the modality-specific generative features along with discriminative representations for classification.

RAVEN. The Recurrent Attended Variation Embedding Network (RAVEN) (Wang et al. 2019) utilizes an attention-based model re-adjusting word embeddings according to auxiliary non-verbal signals.

MulT. The Multimodal Transformer (MulT) (Tsai et al. 2019) extends multimodal transformer architecture with directional pairwise crossmodal attention which translates one modality to another using directional pairwise cross-attention.

MAG-BERT. The Multimodal Adaptation Gate for Bert (MAG-BERT) (Rahman et al. 2020) is an improvement over RAVEN on aligned data with applying multimodal adaptation gate at different layers of the BERT backbone.

MISA. The Modality-Invariant and -Specific Representations (MISA) (Hazarika, Zimmermann, and Poria 2020) incorporate a combination of losses including distributional similarity, orthogonal loss, reconstruction loss and task prediction loss to learn modality-invariant and modality-specific representation.

Model	MAE	Corr	Acc-2	F1-Score
TFN	0.428	0.605	79.86	80.15
LMF	0.431	0.600	79.37	78.65
Human-MM	0.408	0.647	81.32	81.73
Self-MM	0.419	0.616	80.74	80.78

Table 3: Results on SIMS.

Basic Settings

Experimental Details. We use Adam as the optimizer and use the initial learning rate of $5e-5$ for Bert and $1e-3$ for other parameters. For a fair comparison, in our model (Self-MM) and two state-of-the-art methods (MISA and MAG-BERT), we run five times and report the average performance.

Evaluation Metrics. Following the previous works (Hazarika, Zimmermann, and Poria 2020; Rahman et al. 2020), we report our experimental results in two forms: classification and regression. For classification, we report Weighted F1 score (F1-Score) and binary classification accuracy (Acc-2). Specifically, for MOSI and MOSEI datasets, we calculate Acc-2 and F1-Score in two ways: negative / non-negative

Tasks	MSE	Corr	Acc-2	F1-Score
M	0.730	0.781	82.38/83.67	82.48/83.70
M, V	0.732	0.775	82.67/83.52	82.76/83.55
M, A	0.728	0.790	82.80/84.76	82.85/84.75
M, T	0.731	0.789	82.65/84.15	82.66/84.10
M, A, V	0.719	0.789	82.94/84.76	83.05/84.81
M, T, V	0.714	0.797	**84.26**/85.91	84.33/**86.00**
M, T, A	**0.712**	0.797	83.67/85.06	83.72/85.06
M, T, A, V	0.713	**0.798**	84.00/**85.98**	**84.42**/85.95

Table 4: Results for multimodal sentiment analysis with different tasks using Self-MM. M, T, A, V represent the multimodal, text, audio, and vision task, respectively.

(non-exclude zero) (Zadeh et al. 2017) and negative / positive (exclude zero) (Tsai et al. 2019). For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). Except for MAE, higher values denote better performance for all metrics.

Results and Analysis

In this section, we make a detailed analysis and discussion about our experimental results.

Quantitative Results

Table 1 shows the comparative results on MOSI and MOSEI datasets. For a fair comparison, according to the difference of "Data Setting", we split models into two categories: Unaligned and Aligned. Generally, models using aligned corpus can get better results (Tsai et al. 2019). In our experiments, first, comparing with unaligned models (TFN and LMF), we achieve a significant improvement in all evaluation metrics. Even comparing with aligned models, our method gets competitive results. Moreover, we reproduce the two best baselines "MISA" and "MAG-BERT" under the same conditions. We find that our model surpasses them on most of the evaluations.

Since the SIMS dataset only contains unaligned data, we compare the Self-MM with TFN and LMF. Besides, we use the human-annotated unimodal labels to replace the auto-generated u-labels, called Human-MM. Experimental results are shown in Table 3. We can find that the Self-MM gets better results than TFN and LMF and achieve comparable performance with Human-MM. The above results show that our model can be applied to different data scenarios and achieve significant improvements.

Ablation Study

To further explore the contributions of Self-MM, we compare the effectiveness of combining different uni-modal tasks. Results are shown in Table 4. Overall, compared with the single-task model, the introduce of unimodal subtasks can significantly improve model performance. From the results, we can see that "M, T, V" and "M, T, A" achieve comparable or even better results than "M, T, A, V". Moreover, we can find that subtasks, "T" and "A", help more than the subtask "V".

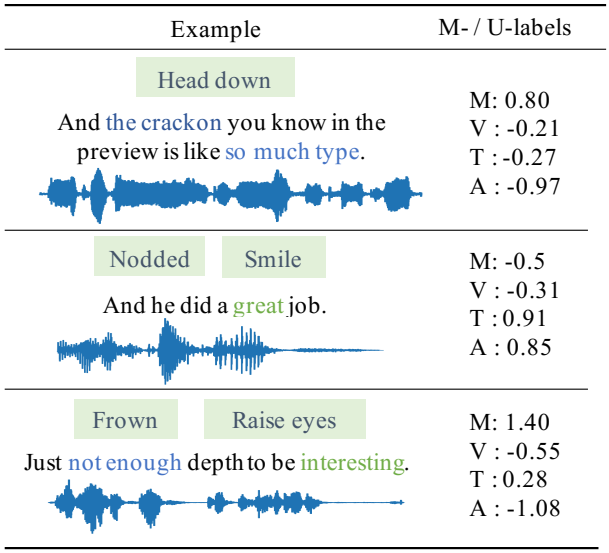

Figure 5: Case study for the Self-MM on MOSI. The "M" is human-annotated, and "V, T, A" are auto-generated.

Case Study

To validate the reliability and reasonability of auto-generated u-labels, we analyze the update process of u-labels, shown in Figure 4. We can see that as the number of iterations increases, the distributions of u-labels tends to stabilize. It is in line with our expectations. Compared with MOSI and SIMS datasets, the update process on the MOSEI has faster convergence. It shows that the larger dataset has more stable class centers, which is more suitable for self-supervised methods.

In order to further show the reasonability of the u-labels, we selected three multimodal examples from the MOSI dataset, as shown in Figure 5. In the first and third cases, human-annotations m-labels are 0.80 and 1.40. However, for single modalities, they are inclined to negative sentiments. In line with expectation, the u-labels get negative offsets on the m-labels. A positive offset effect is achieved in the second case. Therefore, the auto-generated u-labels are significant. We believe that these independent u-labels can aid in learning modality-specific representation.

Conclusion and Future Work

In this paper, we introduce unimodal subtasks to aid in learning modality-specific representations. Different from previous works, we design a unimodal label generation strategy based on the self-supervised method, which saves lots of human costs. Extensive experiments validate the reliability and stability of the auto-generated unimodal labels. We hope this work can provide a new perspective on multimodal representation learning.

We also find that the generated audio and vision labels are not significant enough limited by the pre-processed features. In future work, we will build an end-to-end multimodal learning network and explore the relationship between unimodal and multimodal learning.

Acknowledgments

This paper is supported by seed fund of Tsinghua University (Department of Computer Science and Technology) - Siemens Ltd., China Joint Research Center for Industrial Intelligence and Internet of Things.

References

Akhtar, M. S.; Chauhan, D.; Ghosal, D.; Poria, S.; Ekbal, A.; and Bhattacharyya, P. 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, Volume 1 (Long and Short Papers)*, 370–379.

Baltrušaitis, T.; Ahuja, C.; and Morency, L. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2): 423–443.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. *CoRR* abs/2005.03545. URL https://arxiv.org/abs/2005.03545.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Liu, W.; Mei, T.; Zhang, Y.; Che, C.; and Luo, J. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3707–3715.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A. B.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256.

Poria, S.; Hazarika, D.; Majumder, N.; and Mihalcea, R. 2020. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *arXiv preprint arXiv:2005.00357* .

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A. B.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2359–2369.

Sun, Z.; Sarma, P.; Sethares, W.; and Liang, Y. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8992–8999.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning Factorized Multimodal Representations. In *International Conference on Learning Representations*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7216–7223.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020a. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3718–3727. Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.343. URL https://www.aclweb.org/anthology/2020.acl-main.343.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020b. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3718–3727.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* .

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927* .

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* .

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

Zhang, W.; Li, R.; Zeng, T.; Sun, Q.; Kumar, S.; Ye, J.; and Ji, S. 2016. Deep model based transfer and multi-task learning for biological image analysis. *IEEE transactions on Big Data* .

Zhang, Y.; and Yang, Q. 2017. A Survey on Multi-Task Learning. *CoRR* abs/1707.08114. URL http://arxiv.org/abs/1707.08114.