CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotations of Modality

Wenmeng Yu[1], **Hua Xu**[2]*, **Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, Kaicheng Yang**
State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology
ywm18@mails.tsinghua.edu.cn[1], xuhua@tsinghua.edu.cn[2]

Abstract

Previous studies in multimodal sentiment analysis have used limited datasets, which only contain unified multimodal annotations. However, the unified annotations do not always reflect the independent sentiment of single modalities and limit the model to capture the difference between modalities. In this paper, we introduce a Chinese single- and multi-modal sentiment analysis dataset, CH-SIMS, which contains 2,281 refined video segments in the wild with both multimodal and independent unimodal annotations. It allows researchers to study the interaction between modalities or use independent unimodal annotations for unimodal sentiment analysis. Furthermore, we propose a multi-task learning framework based on late fusion as the baseline. Extensive experiments on the CH-SIMS show that our methods achieve state-of-the-art performance and learn more distinctive unimodal representations. The full dataset and codes are available for use at https://github.com/thuiar/MMSA.

1 Introduction

Sentiment analysis is an important research area in Natural Language Processing (NLP). It has wide applications for other NLP tasks, such as opinion mining, dialogue generation, and user behavior analysis. Previous study (Pang et al., 2008; Liu and Zhang, 2012) mainly focused on text sentiment analysis and achieved impressive results. However, using text alone is not sufficient to determine the speaker's sentimental state, and text can be misleading. With the booming of short video applications, nonverbal behaviors (vision and audio) are introduced to solve the above shortcomings (Zadeh et al., 2016; Poria et al., 2017).

In multimodal sentiment analysis, intra-modal representation and inter-modal fusion are two important and challenging subtasks (Baltrušaitis et al., 2018; Guo et al., 2019). For intra-modal representation, it is essential to consider the temporal or spatial characteristics in different modalities. The methods based on Convolutional Neural Network (CNN), Long Short-term Memory (LSTM) network and Deep Neural Network (DNN) are three representative approaches to extract unimodal features (Cambria et al., 2017; Zadeh et al., 2017, 2018a). For inter-modal fusion, numerous methods have been proposed in recent years. For example, concatenation (Cambria et al., 2017), Tensor Fusion Network (TFN) (Zadeh et al., 2017), Low-rank Multimodal Fusion (LMF) (Liu et al., 2018),

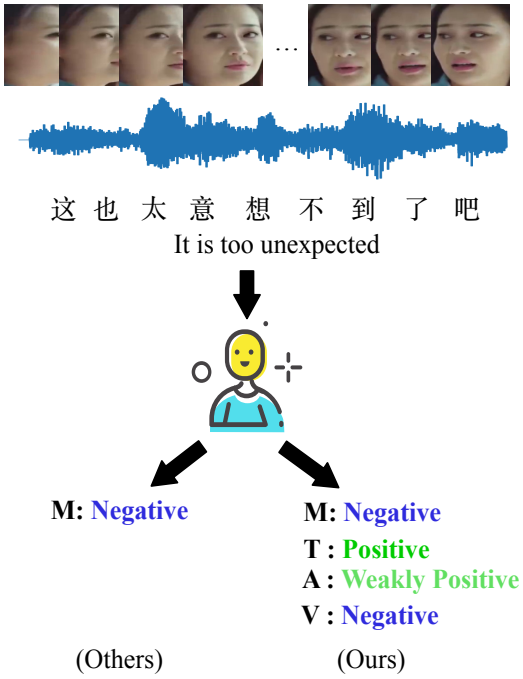

Figure 1: An example of the annotation difference between CH-SIMS and other datasets. For each multimodal clip, in addition to multimodal annotations, our proposed dataset has independent unimodal annotations. M: Multimodal, T: Text, A: Audio, V: Vision.

*Corresponding Author

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3718–3727

Memory Fusion Network (MFN) (Zadeh et al., 2018a), Dynamic Fusion Graph (DFG) (Zadeh et al., 2018b), and others. In this paper, we mainly consider late-fusion methods that perform intra-modal representation learning first and then employ inter-modal fusion. An intuitive idea is that the greater the difference between inter-modal representations, the better the complementarity of inter-modal fusion. However, it is not easy for existing late-fusion models to learn the differences between different modalities, further limits the performance of fusion. The reason is that the existing multimodal sentiment datasets only contain a unified multimodal annotation for each multimodal segment, which is not always suitable for all modalities. In other words, all modalities share a standard annotation during intra-modal representation learning. Further, these unified supervisions will guide intra-modal representations to be more consistent and less distinctive.

To validate the above analysis, in this paper, we propose a Chinese multimodal sentiment analysis dataset with independent unimodal annotations, CH-SIMS. Figure 1 shows an example of the annotation difference between our proposed dataset and the other existing multimodal datasets. SIMS has 2,281 refined video clips collected from different movies, TV serials, and variety shows with spontaneous expressions, various head poses, occlusions, and illuminations. The CHEAVD (Li et al., 2017) is also a Chinese multimodal dataset, but it only contains two modalities (vision and audio) and one unified annotation. In contrast, SIMS has three modalities and unimodal annotations except for multimodal annotations for each clip. Therefore, researchers can use SIMS to do both unimodal and multimodal sentiment analysis tasks. Furthermore, researchers can develop new methods for multimodal sentiment analysis with these additional annotations.

Based on SIMS, we propose a multimodal multi-task learning framework using unimodal and multimodal annotations. In this framework, the unimodal and multimodal tasks share the feature representation sub-network in the bottom. It is suitable for all multimodal models based on late-fusion. Then, we introduce three late-fusion models, including TFN, LMF, and Late-Fusion DNN (LF-DNN), into our framework. With unimodal tasks, the performance of multimodal task is significantly increased. Furthermore, we make a detailed discussion on multimodal sentiment analysis, unimodal sentiment analysis and multi-task learning. Lastly, we verify that the introduction of unimodal annotations can effectively expand the difference between different modalities and obtain better performance in inter-modal fusion.

In this work, we provide a new perspective for multimodal sentiment analysis. Our main contributions in this paper can be summarized as follows:

- We propose a Chinese multimodal sentiment analysis dataset with more fine-grained annotations of modality, CH-SIMS. These additional annotations make our dataset available for both unimodal and multimodal sentiment analysis.
- We propose a multimodal multi-task learning framework, which is suitable for all late-fusion methods in multimodal sentiment analysis. Besides, we introduce three late-fusion models into this framework as strong baselines for SIMS.
- The benchmark experiments on the SIMS show that our methods learn more distinctive unimodal representations and achieve state-of-the-art performance.

2 Related Work

In this section, we briefly review related work in multimodal datasets, multimodal sentiment analysis, and multi-task learning.

2.1 Multimodal Datasets

To meet the needs of multimodal sentiment analysis and emotion recognition, researchers have proposed various of multimodal datasets, including IEMOCAP (Busso et al., 2008), YouTube (Morency et al., 2011), MOUD (Pérez-Rosas et al., 2013), ICT-MMMO (Wöllmer et al., 2013), MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018b) and so on. In addition, Li et al. (2017) proposed a Chinese emotional audio-visual dataset and Poria et al. (2018) proposed a multi-party emotional, conversational dataset containing more than two speakers per dialogue. However, these existing multimodal datasets only contain a unified multimodal annotation for each multimodal corpus. In contrast, SIMS contains both unimodal and multimodal annotations.

Item	#
Total number of videos	60
Total number of segments	2,281
- Male	1,500
- Female	781
Total number of distinct speakers	474
Average length of segments (s)	3.67
Average word count per segments	15

Table 1: Statistics of SIMS Dataset.

2.2 Multimodal Sentiment Analysis

Multimodal sentiment analysis has become a major research topic that integrates verbal and nonverbal behaviors. Cambria et al. (2017) proposed a general multimodal sentiment analysis framework that is composed of representation learning on intra-modality and feature concatenation on inter-modality. Based on this framework, many studies focused on designing a new fusion network to capture better multimodal representations and achieve better performance. Zadeh et al. (2017) proposed a tensor fusion network, which obtains a new tensor representation by computing the outer product between unimodal representations. Liu et al. (2018) used a low-rank multimodal fusion method to decompose the weight tensor and decrease the computational complexity of tensor-based methods. Zadeh et al. (2018a) designed a memory fusion network with a special attention mechanism for cross-view interactions. Tsai et al. (2019) proposed crossmodal transformers to reinforce a target modality from another source modality by learning the attention across the two modalities' features. Tsai et al. (2018) learned meaningful multimodal representations by factorizing representations into two sets of independent factors: multimodal discriminative and modality-specific generative factors. Different from the above methods, we aim to learn more distinctive unimodal representations by introducing independent unimodal annotations.

2.3 Multi-task Learning

Multi-task learning aims to improve the generalization performance of multiple related tasks by utilizing useful information contained in these tasks (Zhang and Yang, 2017). A classical method is that different tasks share the first several layers and then have task-specific parameters in the subsequent layers (Liu et al., 2015; Zhang et al., 2016b). Based on this method, we design a multimodal multi-task learning framework for verifying the practicality and feasibility of independent unimodal annotations.

3 CH-SIMS Dataset

In this section, we introduce a novel Chinese multimodal sentiment analysis dataset with independent unimodal annotations, CH-SIMS. In the following subsections, we will explain the data acquisition, annotation, and feature extraction in detail.

3.1 Data Collection

Comparing with unimodal datasets, the requirements of multimodal datasets are relatively high. A fundamental requirement is that the speaker's face and voice must appear in the picture at the same time and remain for a specific period of time. In this work, to acquire video clips as close to life as possible, we collect target fragments from movies, TV series, and variety shows. After getting raw videos, we use video editing tools, Adobe Premiere Pro[1], to crop target segments at the frame level, which is very time-consuming but accurate enough. Moreover, during the data collection and cropping, we enforce the following constraints:

- We only consider mandarin and are cautious with the selection of materials with the accent.
- The length of clips is no less than one second and no more than ten seconds.
- For each video clip, no other faces appear except for the speaker's face.

Finally, we collect 60 raw videos and acquire 2,281 video segments. SIMS has rich character background, wide age range, and high quality. Table 1 shows the basic statistics for SIMS.[2]

3.2 Annotation

We make one multimodal annotation and three unimodal annotations for each video clip. In addition to the increase in workload, the mutual interference between different modalities is more confused. To avoid this problem as much as possible, we claim every labeler can only see the information in the current modality when annotating. Besides, conducting four annotations at the same time is not

[1] https://www.adobe.com/products/premiere.html

[2] We consulted a legal office to verify that the academic usage and distribution of very short length videos fall under the fair use category.

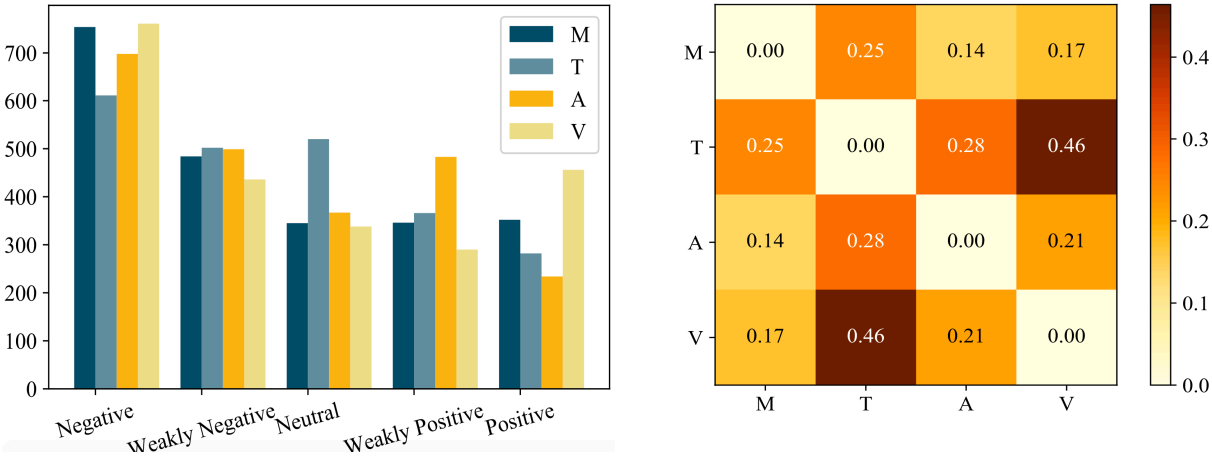

Figure 2: Left: the distribution of sentiment over the entire dataset in one **M**ultimodal annotation and three single-modal (**T**ext, **A**udio, and **V**ision) annotations. Right: the confusion matrix shows the annotations difference between different modalities in CH-SIMS. The larger the value, the greater the difference.

permitted. More precisely, every labeler makes unimodal annotation first and then performs multimodal annotation, which of the order is text first, audio second, then silent video, and multimodal last.

For each clip, every annotator decides its sentimental state as -1 (negative), 0 (neutral) or 1 (positive). we have five independent students in this field making annotations. Then, in order to do both regression and multi-classifications tasks, we average the five labeled results. Therefore, the final labeling results are one of {-1.0, -0.8, -0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8, 1.0}. We further divide these values into 5 classifications: negative {-1.0, -0.8}, weakly negative {-0.6, -0.4, -0.2}, neutral {0.0}, weakly positive {0.2, 0.4, 0.6} and positive {0.8, 1.0}.

The histogram in the left of Figure 2 shows the distribution of sentiment over the entire dataset in four annotations. We can see that negative segments are more than positive segments. The main reason is that actors in film and television dramas are more expressive in negative sentiments than positive ones. The confusion matrix in the right of Figure 2 indicates the annotations difference between different modalities, which is computed as:

$$D_{ij} = \frac{1}{N}\sum_{n=1}^{N}(A_i^n - A_j^n)^2 \qquad (1)$$

where $i, j \in \{m, t, a, v\}$, N is the number of all samples, A_i^n means the n_{th} label value in modal i.

From the confusion matrix, we can see that the difference between A and M is minimal, and the difference between V and T is maximal, which is in line with expectations. Because audio contains text information, closer to multimodal while the connection between video and text is sparse.

Furthermore, we provide the other attribute annotations, including speakers' age and gender. And we use sentimental annotations only in our following experiments.

3.3 Extracted Features

The extracted features for all modalities are as follows (we use the same basic features in all experiments):

Text: All videos have manual transcription, including the Chinese and English versions. We use Chinese transcriptions only. We add two unique tokens to indicate the beginning and the end for each transcript. And then, pre-trained Chinese BERT-base word embeddings are used to obtain word vectors from transcripts (Devlin et al., 2018). It is worth noting that we do not use word segmentation tools due to the characteristic of BERT. Eventually, each word is represented as a 768-dimensional word vector.

Audio: We use LibROSA (McFee et al., 2015) speech toolkit with default parameters to extract acoustic features at 22050Hz. Totally, 33-dimensional frame-level acoustic features are extracted, including 1-dimensional logarithmic fundamental frequency (log F0), 20-dimensional Mel-frequency cepstral coefficients (MFCCs) and 12-dimensional Constant-Q chromatogram (CQT). These features are related to emotions and tone of speech according to (Li et al., 2018).

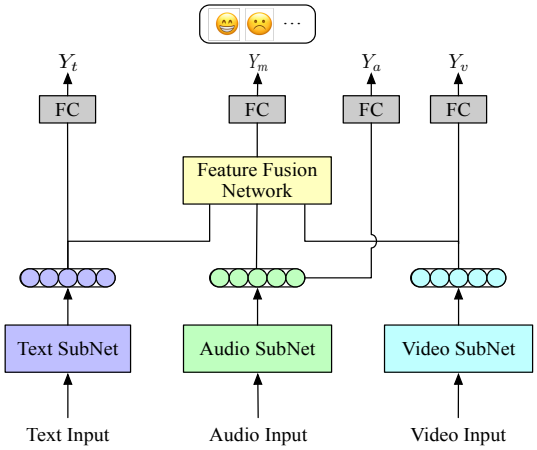

Figure 3: Multimodal multi-task learning framework.

Vision: Frames are extracted from the video segments at 30Hz. We use the MTCNN face detection algorithm (Zhang et al., 2016a) to extract aligned faces. Then, following Zadeh et al. (2018b), we use MultiComp OpenFace2.0 toolkit (Baltrusaitis et al., 2018) to extract the set of 68 facial landmarks, 17 facial action units, head pose, head orientation, and eye gaze. Lastly, 709-dimensional frame-level visual features are extracted in total.

4 Multimodal Multi-task Learning Framework

In this section, we describe our proposed multimodal multi-task learning framework. Shown as Figure 3, based on late-fusion multimodal learning framework (Cambria et al., 2017; Zadeh et al., 2017), we add independent output units for three unimodal representations: text, audio, and vision. Therefore, these unimodal representations not only participate in feature fusion but are used to generate their predictive outputs.

For the convenience in following introduction, in **t**ext, **a**udio and **v**ision, we assume that L^u, D_i^u, D_r^u, where $u \in \{t, a, v\}$, represent the sequence length, initial feature dimension extracted by section 3.3 and representation dimension learned by unimodal feature extractor, respectively. The batch size is B.

4.1 Unimodal SubNets

Unimodal subNets aim to learn intra-modal representations from initial feature sequences. A universal feature extractor can be formalized as:

$$R_u = S_u(I_u) \tag{2}$$

where $I_u \in R^{B \times L^u \times D_i^u}$, $R_u \in R^{B \times D_r^u}$. $S_u(\bullet)$ is the feature extractor network for modal u.

In this work, following Zadeh et al. (2017); Liu et al. (2018), we use a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network, a deep neural network with three hidden layers of weights W_a and a deep neural network with three hidden layers of weights W_v to extract textual, acoustic and visual embeddings, respectively.

4.2 Feature Fusion Network

Feature fusion network aims to learn inter-modal representation with three unimodal representations, formulated as:

$$R_m = F(R_t, R_a, R_v) \tag{3}$$

where $R_t, R_a, R_v \in R^{B \times D_r^u}$ are the unimodal representations. $F(\bullet)$ is the feature fusion network and R_m is the fusion representation.

In this work, for full comparison with existing works, we try three fusion methods: LF-DNN, TFN (Zadeh et al., 2017) and LMF (Liu et al., 2018).

4.3 Optimization Objectives

Except for the training losses in different tasks, we sparse the sharing parameters via L2 norm, which aims to select intra-modal features. Therefore, our optimization objectives is:

$$\min \frac{1}{N_t} \sum_{n=1}^{N_t} \sum_i \alpha_i L(y_i^n, \hat{y}_i^n) + \sum_j \beta_j ||W_j||_2^2 \tag{4}$$

where N_t is the number of training samples, $i \in \{m, t, a, v\}$, $j \in \{t, a, v\}$. $L(y_i^n, \hat{y}_i^n)$ means the training loss of n_{th} sample in modality i. W_j is the sharing parameters in modality j and multimodal tasks. α_i is the hyperparameter to balance different tasks and β_j represents the step of weight decay of subNet j, respectively.

Lastly, we use a three-layer DNN to generate outputs of different tasks. In this work, we treat these tasks as regression models and use the L1 loss as training loss in Equation 4.

5 Experiments

In this section, we mainly explore the following problems using SIMS:

(1) Multimodal Sentiment Analysis: We evaluate the performance of multimodal multi-task learning methods comparing with the other methods. The aim is to validate the advantages of multi-task

Model	Acc-2	Acc-3	Acc-5	F1	MAE	Corr
EF-LSTM	69.37 ± 0.0	51.73 ± 2.0	21.02 ± 0.2	81.91 ± 0.0	59.34 ± 0.3	-04.39 ± 2.8
MFN	77.86 ± 0.4	63.89 ± 1.9	39.39 ± 1.8	78.22 ± 0.4	45.19 ± 1.2	55.18 ± 2.0
MULT	77.94 ± 0.9	65.03 ± 2.1	35.34 ± 2.9	79.10 ± 0.9	48.45 ± 2.6	55.94 ± 0.6
LF-DNN	79.87 ± 0.6	66.91 ± 1.2	**41.62 ± 1.4**	80.20 ± 0.6	42.01 ± 0.9	61.23 ± 1.8
MLF-DNN*	82.28 ± 1.3	**69.06 ± 3.1**	38.03 ± 6.0	82.52 ± 1.3	**40.64 ± 2.0**	**67.47 ± 1.8**
▽	**↑ 2.41**	**↑ 2.15**	↓ 3.59	**↑ 2.32**	**↓ 1.37**	**↑ 6.24**
LMF	79.34 ± 0.4	64.38 ± 2.1	35.14 ± 4.6	79.96 ± 0.6	43.99 ± 1.6	60.00 ± 1.3
MLMF*	82.32 ± 0.5	67.70 ± 2.2	37.33 ± 2.5	**82.66 ± 0.7**	42.03 ± 0.9	63.13 ± 1.9
▽	**↑ 2.98**	**↑ 3.32**	**↑ 2.19**	**↑ 2.70**	**↓ 1.96**	**↑ 3.13**
TFN	80.66 ± 1.4	64.46 ± 1.7	38.38 ± 3.6	81.62 ± 1.1	42.52 ± 1.1	61.18 ± 1.2
MTFN*	**82.45 ± 1.3**	69.02 ± 0.3	37.20 ± 1.8	82.56 ± 1.2	40.66 ± 1.1	66.98 ± 1.3
▽	**↑ 1.79**	**↑ 4.56**	↓ 1.18	**↑ 0.94**	**↓ 1.86**	**↑ 5.80**

Table 2: (%) Results for sentiment analysis on the CH-SIMS dataset. The models with $*$ are multi-task models, extended from single-task models by introducing independent unimodal annotations. For example, MLF-DNN* is the extension of LF-DNN. The rows with ▽ means the improvements or reductions of new models compared to original ones in the current evaluation metric.

learning with unimodal annotations and set up multimodal baselines for SIMS.

(2) Unimodal Sentiment Analysis: We analyze the performance in unimodal tasks with unimodal or multimodal annotations only. The aim is to validate the necessary of multimodal analysis and set unimodal baselines for SIMS.

(3) Representations Differences: We use t-SNE to visualize the unimodal representations of models with or without independent unimodal annotations. The aim is to show that the learned unimodal representations are more distinctive after using unimodal annotations.

Item	Total	NG	WN	NU	WP	PS
#Train	1,368	452	290	207	208	211
#Valid	456	151	97	69	69	70
#Test	457	151	97	69	69	71

Table 3: Dataset splits in SIMS. We split train, valid and test set in 6:2:2. NG: Negative, WN: Weakly Negative, NU: Neutral, WP: Weakly Positive, PS: Positive.

5.1 Baselines

In this section, we briefly review our baselines used in the following experiments.

Early Fusion LSTM. The Early Fusion LSTM (EF-LSTM) (Williams et al., 2018) concatenates initial inputs of three modalities first and then use LSTM to capture long-distance dependencies in a sequence.

Later Fusion DNN. In contrast with EF-LSTM, the Later Fusion DNN (LF-DNN) learns unimodal features first and then concatenates these features before classification.

Memory Fusion Network. The Memory Fusion Network (MFN) (Zadeh et al., 2018a) accounts for view-specific and cross-view interactions and continuously models them through time with a special attention mechanism and summarized through time with a Multi-view Gated Memory. MFN needs word-level alignment in three modalities. However, this is not easy for SIMS because we haven't found a reliable alignment tool of Chinese corpus. In this work, we follow Tsai et al. (2019) to use CTC (Graves et al., 2006) as an alternative.

Low-rank Multimodal Fusion. The Low-rank Multimodal Fusion (LMF) (Liu et al., 2018) model learns both modality-specific and cross-modal interactions by performing efficient multimodal fusion with modality-specific low-rank factors.

Tensor Fusion Network. The Tensor Fusion Network (TFN) (Zadeh et al., 2017) explicitly models view-specific and cross-view dynamics by creating a multi-dimensional tensor that captures unimodal, bimodal and trimodal interactions across three modalities.

Multimodal Transformer. The Multimodal Transformer (MULT) (Tsai et al., 2019) using the directional pairwise crossmodal attention to realize the interactions between multimodal sequences across distinct time steps and latently adapt streams from one modality to another.

Task	Label	Acc-2	F1	MAE	Corr
A	A	67.70	79.61	53.80	10.07
	M	65.47	71.44	57.89	14.54
V	V	81.62	82.73	49.57	57.61
	M	74.44	79.55	54.46	38.76
T	T	80.26	82.93	41.79	49.33
	M	75.19	78.43	52.73	38.55

Table 4: (%) Results for unimodal sentiment analysis on the CH-SIMS dataset using MLF-DNN. The column of "Label" indicates which annotation we use in this task.

5.2 Experimental Details

In this section, we introduce our experimental settings in detail, including dataset splits, hyper-parameters selection, and our evaluation metrics.

Dataset Splits. We shuffle all video clips in random first and then divide train, valid and, test splits by multimodal annotations. The detailed split results are shown in Table 3.

Hyper-parameters Selection. Due to the different sequence lengths in different segments, it is necessary that fixing sequence length for the specific modality. Empirically, we choose the average length plus three times the standard deviation as the maximum length of the sequence. Besides, for all baselines and our methods, we adjust their hyperparameters using grid search with binary classification accuracy. For a fair comparison, in each experiment, we select five same random seeds (1, 12, 123, 1234, and 12345) and report the average performance of five times.

Evaluation Metrics. The same as Liu et al. (2018); Zadeh et al. (2018b), we record our experimental results in two forms: multi-class classification and regression. For multi-class classification, we report Weighted F1 score and multi-class accuracy Acc-k, where $k \in \{2, 3, 5\}$. For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). Except for MAE, higher values denote better performance for all metrics.

5.3 Results and Discussion

In this section, we present and discuss the experimental results of the research questions introduced in Section 5.

5.3.1 Comparison with Baselines.

We compare three new methods with the aforementioned baselines. In this part, we only consider the multimodal evaluation results though new methods are multi-task. Results are shown in Table 2. Compared with single-task models, multi-task models have better performance in most of evaluation metrics. In particular, all three improved models (MLF-DNN, MLFM, and MTFN) have promotion significantly compared to corresponding original models (LF-DNN, LFM, and TFN) in all evaluation metrics except for Acc-5. The above results demonstrate that the introduction of independent unimodal annotations in multimodal sentiment analysis can significantly improve the performance of existing methods. Also, we find that some methods, such as MULT, that perform well on existing public datasets while they are not satisfactory on SIMS. It further illustrates that designing a robust, cross-lingual multimodal sentiment analysis model is still a challenging task, which is also one of our motivations for proposing this dataset.

5.3.2 Unimodal Sentiment Analysis.

Due to the independent unimodal annotations in SIMS, we conducted two sets of experiments for unimodal sentiment analysis. In the first set of experiments, we use real unimodal labels to verify the model's ability of performing unimodal sentiment analysis. In the second set of experiments, we use multimodal labels instead of unimodal labels to verify the ability of predicting the true emotions of speakers when there is only unimodal information.

Results are shown in Table 4. Firstly, in the same unimodal task, the results under unimodal labels are better than those under multimodal labels. But the former cannot reflect the actual sentimental state of speakers. Secondly, under multimodal annotations, the performance with unimodal information only is lower than using multimodal information in Table 2. Hence, it is inadequate to perform sentiment analysis using unimodal information only due to the inherent limitations of unimodal information.

5.3.3 Representations Differences.

Another motivation for us to propose CH-SIMS is that we think the unimodal representation differences will be greater with independent unimodal annotations. We use t-SNE (Maaten and Hinton, 2008) to visualize intra-modal representations learned in original models (LF-DNN, TFN, and LMF) and new models (MLF-DNN, MTFN, and MLMF), shown as Figure 4. It is relatively obvious that new models learn more distinctive unimodal

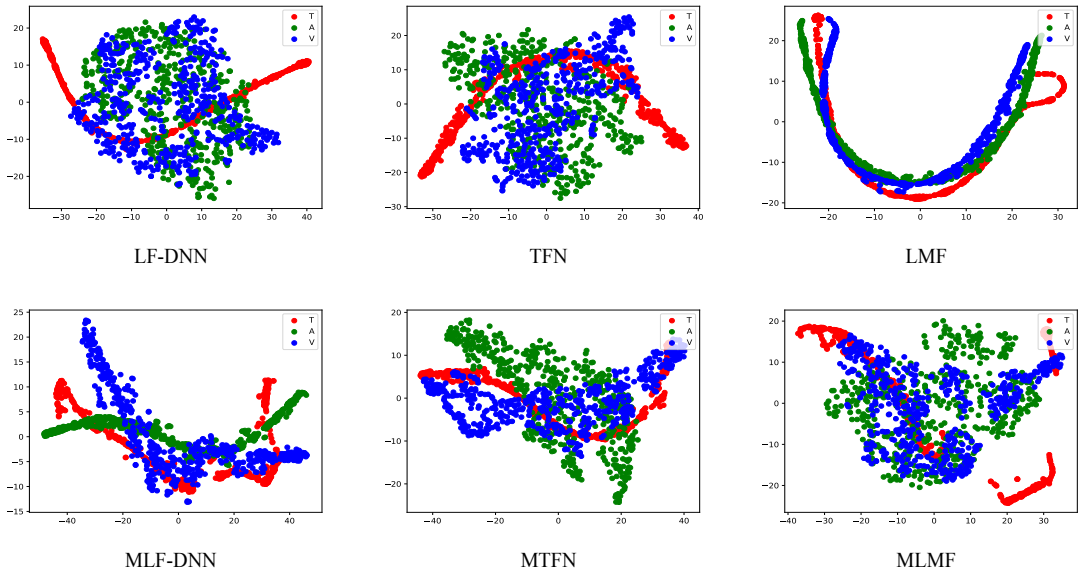

Figure 4: Visualization in Unimodal Representations. In each subfigure, red, green, and blue points represent the unimodal representations in text, audio, and video, respectively. The first row shows the learned representations from models with the multimodal task only. The second row shows the learned representations from multi-task models. The two subgraphs in the same column contrast each other

representations compare to original models. Therefore, unimodal annotations can help the model to obtain more differentiated information and improve the complementarity between modalities.

6 Ablation Study

In this section, we compare the difference in the effects of combining different unimodal tasks on multimodal sentiment analysis. We aim to further explore the influence on multimodal sentiment analysis with different unimodal tasks. Furthermore, we reveal the relationship between multi-task learning and multimodal sentiment analysis.

We conducted multiple combination experiments to analyze the effects of different unimodal subtasks on the main multimodal task. In this part, we only report the results in MLF-DNN. Results are shown in Table 5. The results show that in the case of partial absence of three unimodal subtasks, the performance of the multimodal task has not significantly improved, or even damaged. Two factors may cause an adverse effect in multimodal learning, including the consistency between different unimodal representations and the asynchrony of learning in different tasks. The former means that unified annotations guide the representations to be similar and lack complementarity in different modalities. The latter means that the learning process in different tasks is inconsistent. Taken tasks

Tasks	Acc-2	F1	MAE	Corr
M	80.04	80.40	43.95	61.78
M, T	80.04	80.25	43.11	63.34
M, A	76.85	77.28	46.98	55.16
M, V	79.96	80.38	43.16	61.87
M, T, A	80.88	81.10	42.54	64.16
M, T, V	80.04	80.87	42.42	60.66
M, A, V	79.87	80.32	43.06	62.95
M, T, A, V	**82.28**	**82.52**	**40.64**	**64.74**

Table 5: (%) Results for multimodal sentiment analysis with different tasks using MLF-DNN. "M" is the main task and "T, A, V" are auxiliary tasks. Only the results of task "M" are reported.

"M, A" as an example, the sub-network of subtask "A" is supervised by multimodal loss and unimodal loss. In contrast, subtask "T" and subtask "V" are supervised by their unimodal loss only. It means the "A" is learned twice while the "T" and the "V" are learned once only during an training epoch. Therefore, the introduction of unimodal tasks will reduce the consistency of the representation and strengthen the complementarity, but will also cause the asynchrony. As more unimodal tasks are introduced, the positive effects of the former gradually increase, and the negative effects of the latter gradually decrease. Finally, when all unimodal tasks are added, the negative effect of the latter is almost dis-

appearing. Finally, the performance of the model with tasks "M, T, A, V" reaches a peak.

7 Conclusion

In this paper, we propose a novel Chinese multimodal sentiment analysis dataset with independent unimodal annotations and a multimodal multi-task learning framework based on late-fusion methods. We hope that the introduction of CH-SIMS will provide a new perspective for researches on multimodal analysis. Furthermore, we conduct extensive experiments on discussing unimodal, multimodal, and multi-task learning. Lastly, we summarize our overall findings as follows:

- Multimodal labels cannot reflect unimodal sentimental states always. The unified multimodal annotations may mislead the model to learn inherent characteristics of unimodal representations.
- With the help of unimodal annotations, models can learn more differentiated information and improve the complementarity between modalities.
- When performing multi-task learning, the asynchrony of learning in different subtasks may cause an adverse effect on multimodal sentiment analysis.

In the future, we will further explore the connection between multimodal analysis and multi-task learning and incorporate more fusion strategy, including early- and middle-fusion.

Acknowledgments

This paper is founded by National Natural Science Foundation of China (Grant No: 61673235) and National Key R&D Program Projects of China (Grant No: 2018YFC1707605). We would like to thank the anonymous reviewers for their valuable suggestions.

References

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.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

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. *Language resources and evaluation*, 42(4):335.

Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and RBV Subramanyam. 2017. Benchmarking multimodal sentiment analysis. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 166–179. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

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

Runnan Li, Zhiyong Wu, Jia Jia, Jingbei Li, Wei Chen, and Helen Meng. 2018. Inferring user emotive state changes in realistic human-computer conversational dialogs. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 136–144. ACM.

Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. 2017. Cheavd: a chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8(6):913–924.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3715.

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. *arXiv preprint arXiv:1806.00064*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.

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*, pages 169–176. ACM.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, 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)*, pages 1033–1038. IEEE.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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)*, pages 2236–2246.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016a. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Wenlu Zhang, Rongjian Li, Tao Zeng, Qian Sun, Sudhir Kumar, Jieping Ye, and Shuiwang Ji. 2016b. Deep model based transfer and multi-task learning for biological image analysis. *IEEE transactions on Big Data*.

Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.