Multimodal Sentiment Analysis with Multi-perspective Fusion Network Focusing on Sense Attentive Language

Xia Li[1,2(✉)] and Minping Chen[2]

[1] Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangzhou, China
[2] School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China
{xiali,minpingchen}@gdufs.edu.cn

Abstract. Multimodal sentiment analysis aims to learn a joint representation of multiple features. As demonstrated by previous studies, it is shown that the language modality may contain more semantic information than that of other modalities. Based on this observation, we propose a Multi-perspective Fusion Network(MPFN) focusing on Sense Attentive Language for multimodal sentiment analysis. Different from previous studies, we use the language modality as the main part of the final joint representation, and propose a multi-stage and uni-stage fusion strategy to get the fusion representation of the multiple modalities to assist the final language-dominated multimodal representation. In our model, a Sense-Level Attention Network is proposed to dynamically learn the word representation which is guided by the fusion of the multiple modalities. As in turn, the learned language representation can also help the multi-stage and uni-stage fusion of the different modalities. In this way, the model can jointly learn a well integrated final representation focusing on the language and the interactions between the multiple modalities both on multi-stage and uni-stage. Several experiments are carried on the CMU-MOSI, the CMU-MOSEI and the YouTube public datasets. The experiments show that our model performs better or competitive results compared with the baseline models.

Keywords: Multimodal sentiment analysis · Multimodal fusion · Sense Attentive Language

1 Introduction

Multimodal sentiment analysis is a task of predicting sentiment of a video, an image or a text based on multiple modal features. With the increase of short videos on the internet, such as Douyin, YouTube, etc., multimoal sentiment analysis can be used to analyze the opinions of the public based on the speaker's language, facial gestures and acoustic behaviors.

M. Sun et al. (Eds.): CCL 2020, LNAI 12522, pp. 359–373, 2020.
https://doi.org/10.1007/978-3-030-63031-7_26

Based on the successes in video, image, audio and language processing, multimodal sentiment analysis has been studied extensively and produced impressive results in recent years [6–8,20,25]. The core of the multimodal sentiment analysis is to capture a better fusion of different modalities. Different methods are proposed to fuse the multimodal features and help to capture the interactions of the modalities. Tensor Fusion Network [22] is proposed to obtain raw unimodal representations, bimodal interactions and tri-modal interactions in the form of 2D-tensor and 3D-tensor simultaneously. Low-rank Fusion Network [7] is then proposed to alleviate the drawback of the large amount of parameters by low-rank factor. Although the above methods achieved good results, they treat all modalities equally and fuse the modalities in the same contribution. We find that language modality always contain more semantic information for sentiment analysis, that's why most of ablation experiments of previous studies [8,15,22] show that when using features from only one modality, the model using language features performs much better than using vision features or acoustic features.

In this paper, we take the assumption that the language modality contains more information than that of the vision and acoustic modalities. We regard language as the major modality and hope to use other modalities to assist the language modality to produce better performance for multimodal sentiment analysis. To this end, we propose a multi-perspective fusion network for multimodal sentiment analysis focusing on sense attentive language. Our model focuses on two aspects: (1) getting rich semantic language representation through the fusion of the sense level attention of language guided by other modalities. (2) learning comprehensive multimodal fusion from multiple perspectives, as well as keeping the enhanced language representation.

In order to get rich semantic information of the language modality, we incorporate a sense-level attention network into the model to obtain a more elaborate representation of the language. Generally speaking, there are many words which have more than one sense and their different senses may lead to different sentiment of a text in different context. Previous studies try to distinguish the ambiguities of a word from the text modality [21,27] using HowNet [4] and LIWC [12], while we hope the sense of a word can be distinguished not only by the context of the text but also by fusion of other modalities(video and acoustic). As an example shown in Fig. 1, we hope to predict the sentiment of the language "It would make sense". As can be seen, the word "sense" in the language modality has a higher attention weight which could be guided by the "smile face of the vision modality" and "high sound audio modality", and also by the "common sense" of the word "sense", which expresses more positive sentiment.

For the effectiveness of modal fusion, the key problem is to model the gap between different modalities and to learn a better multimodal fusion. In this paper, we propose a multi-stage and uni-stage strategy to fuse the multiple modalities in order to capture the interactions between multi-stage sharing information and global information integrated from uni-stage fusion. For multi-stage fusion, we use CNN with different window sizes to capture the multimodal fusion of consecutive temporals within different windows respectively. As for uni-stage

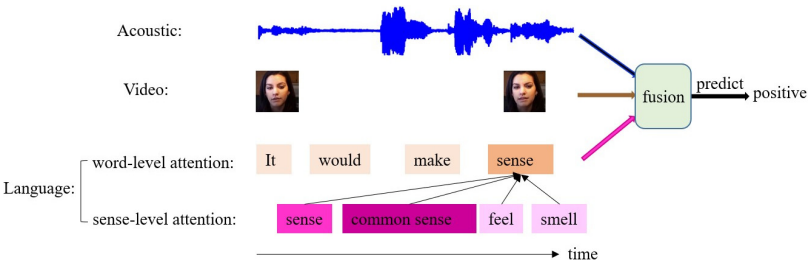

Fig. 1. The sense-level attention and word-level attention of the text "It would make sense" learned by our model. The first line is the acoustic modality, the second line is the video modality. The third line and the last line are the language modality, in which the third line is the original sentence, and the last line presents the senses of word "sense". Darker color means greater weight.

fusion, we first perform a projection operation on the concatenation of the LSTM outputs of three modalities, then attention mechanism is applied to learn the different contributions of the multimodal features at each temporal and produce a summary, which is regarded as the global multimodal fusion. The main contributions of our work are as follows:

1) To the best of our knowledge, this is the first time to use WordNet to reduce ambiguity for the task of multimodal sentiment analysis, which dynamically learn the different weights of the sense words to produce sense-attentive language presentation.
2) We propose to take language as the major modality and learn multimodal fusion from multi-stage and uni-stage perspective. Our final representation not only contains multimodal fusion, but also keeps the language representation, which is helpful in multimodal sentiment analysis.
3) Our model outperforms the baseline models on the CMU-MOSI, the CMU-MOSEI and the YouTube datasets and the ablation study shows the effectiveness of each components in our model.

2 Related Work

Compared with conventional text-based sentiment analysis, sentiment analysis with multiple modalities achieves significant improvements [1]. One of the most challenging task in multimodal sentiment analysis is to learn a joint representation of multiple modalities.

Earlier work uses fusion approaches such as concatenation of multi-modality features [5,11], while recent studies propose more sophisticated fusion approaches. Poria et al. [15] propose a LSTM-based model to capture contextual information. Zadeh et al. [22] propose a Tensor Fusion Network to explicitly aggregate unimodal, bimodal and trimodal interactions. Liu et al. [7] propose a Low-rank Fusion Network to alleviate the drawback of the large amount of

parameters by low-rank factor. Chen et al. [2] propose a Gated Multimodal Embedding model to learn an on-off switch to filter noisy or contradictory modalities.

As the modalities can have interactions between different timestamps, several models are proposed to fuse the multiple modals from different views. Zadeh et al. [25] propose a Multi-attention Recurrent Network (MARN) to capture the interaction between modalities at different timestamps. Zadeh et al. [23] propose a Memory Fusion Network to learn view-specfic interactions and use an attention mechanism called the Delta-memory Attention Network (DMAN) to identify the cross-view interactions. Liang et al. [6] propose a Recurrent Multistage Fusion Network (RMFN) to model cross-modal interactions using multi-stage fusion approach, in which each stage of fusion focuses on a different subset of multimodal signals, learning increasingly discriminative multimodal representations.

Recently, Pham et al. [14] propose to learn joint representations based on translations between modalities. They use a cycle consistency loss to ensure that the joint representations retain maximal information from all modalities. Instead of directly fusing features at holistic level, Mai et al. [8] propose a strategy named 'divide, conquer and combine' for multimodal fusion. Their model performs fusion hierarchically to consider both local and global interactions. Wang et al. [20] propose a Recurrent Attended Variation Embedding Network (RAVEN) to model expressive nonverbal representations by analyzing the negrained visual and acoustic patterns. Tsai et al. [19] introduce a model that factorizes representations into two sets of independent factors: multimodal discriminative and modality-specic generative factors to optimize for a joint generative-discriminative objective across multimodal data and labels.

Although previous studies propose many effective approaches, most of them treat all modalities equally during the learning of multimodal fusion, which are different from our approach. In our model, we propose a sense-level attention network to learn different word representation under different senses. With the sense-attentive word representation, we can learn enhanced language representation. In addition, we try to learn sufficient multimodal fusion through multi-stage fusion and uni-stage fusion, as well as keeping the language representation to form our final representation.

3 Our Model

Our model consists of three components: sense attentive language representation which is regarded as the main representation of the multimodal fusion; multi-stage multimodal fusion which is designed to capture the interactions between the sharing information on the multi-stage; uni-stage multimodal fusion which is used to capture the global fusion information. The whole architecture of our model is shown in Fig. 2. In the following sections, we will introduce the sense-level attention network in Sect. 3.1, and describe the multi-stage multimodal fusion and the uni-stage multimodal fusion strategy in Sect. 3.2. Section 3.3 describes the final representation and model training.

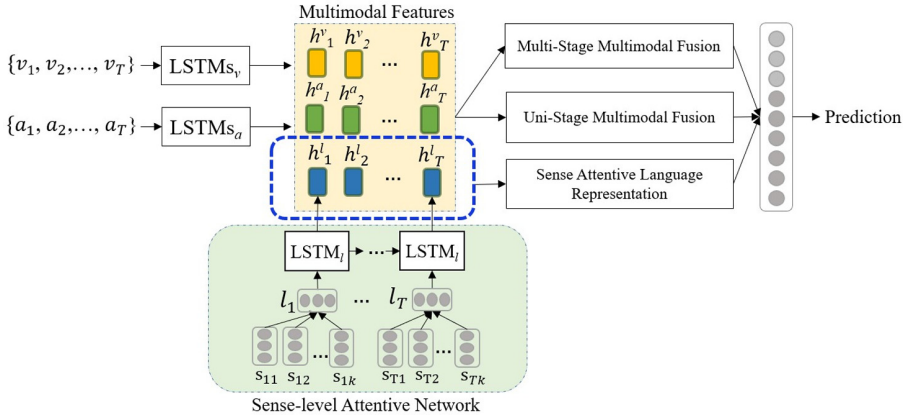

Fig. 2. The whole architecture of our model. The sense-level attention is used to learn the different importance of the sense words of each word in the language modality and produce a sense-attentive representation of language. LSTM layers are then used to model the features from language, vision and acoustic modalities. Three blocks are used to learn multi-stage multimodal fusion, uni-stage multimodal fusion and language representation respectively, which are concatenated to form the final representation.

3.1 Sense-Level Attention Network

As language has rich semantic information, a word may has different senses in different contexts, which may make the sentiment of a sentence totally different. However, the word's embedding representation is unique in the pretrained embeddings. In order to let the model to better distinguish different meanings of a same word, similar to the work of [21,27], we use WordNet to get k number of different senses of a word into the model. If a word don't have any sense in WordNet, we input k number of original words into the model. If there are more than k number of senses for the word, we take the first k number of senses in order and pad the sense sequence with the original word if the number of senses of the word is less than k. We denote the sense sequence of the i-th word in the sentence as $S_i = \{s_{i1}, s_{i2}, \ldots, s_{ik}\}$. The word senses and the original word are converted into embeddings to be input into the model. Then attention mechanism is used to learn the importance weight of different senses of a word and the weighted sum of the embeddings of different senses forms the new representation l_i of the word, as shown in Eqs. (1–3), where W_i and u_i are the trainable weights, b_i is the bias.

$$o_{ij} = relu\left(W_i s_{ij} + b_i\right) \tag{1}$$

$$\alpha_{ij} = softmax\left(u_i o_{ij}\right) \tag{2}$$

$$l_i = \sum_{j=1}^{k} a_{ij} s_{ij} \tag{3}$$

3.2 Multi-stage and Uni-stage Multimodal Fusion

In order to obtain comprehensive multimodal fusion, we propose two strategies to learn the relationship and interactive information between multiple modal features, which are multi-stage fusion and uni-stage fusion. The two strategies are shown in Fig. 3.

After getting the new representation of language modality and the original features of acoustic and vision modality, denoted as $L = \{l_1,\ l_2,\ \ldots, l_T\}, A = \{a_1, a_2,\ .., a_T\}$ and $V = \{v_1, v_2,\ .., v_T\}$ respectively. We use three LSTM layers for modeling the features, aiming to consider the interrelationship of the individual modality in different timestamps. The outputs of LSTM of acoustic, vision and language modality are denoted as $H_A = \{h_1^a,\ h_2^a, \ldots, h_T^a\}, H_V = \{h_1^v,\ h_2^v, \ldots, h_T^v\}$ and $H_L = \{h_1^l,\ h_2^l, \ldots, h_T^l\}$ respectively.

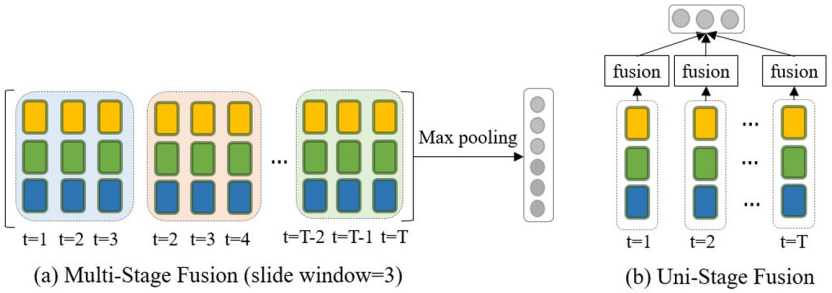

Fig. 3. The strategies of multimodal fusion proposed in our model. The multi-stage fusion aims to capture the interactions of the shared multimodal information in different timestamps. The uni-stage fusion aims to capture the global interactions of multimodal features fused within the same timestamp.

Multi-stage Multimodal Fusion. First we concatenate h_i^a, h_i^v and h_i^l, then we use different CNN layers with different window sizes to learn the multi-stage shared fusion. For CNN with window size 1, we aim to model the relationship between the three modalities timestamp by timestamp, through which we can get the fusion about the word, facial expression and speech tone of the speaker at the same timestamp. For CNN with window size bigger than 1, we aim to model the relationship between the three modalities within several timestamps. We perform maxpooling operation on top of the CNNs respectively and concatenate the results, getting the multi-stage shared multimodal fusion $h_{multi-stage}$. The convolution operation of CNN is shown in Eq. (5–6), where W_z and b_z are trainable weights and bias respectively, w is the window size, f is activation

function which is $relu$ in our implementation and [ ] denotes for concatenation.

$$h_i = \left[h_i^a, h_i^v, h_i^l\right] \tag{4}$$

$$z_i = f\left(W_z\left[h_i : h_{i+w-1}\right] + b_z\right) \tag{5}$$

$$Z_w = maxpooling\left(\left[z_1, z_2, ..., z_T\right]\right) \tag{6}$$

As stated above, we use different CNN layers with different window sizes following maxpooling operation, getting Z_w representation ($w = 1, 2, ...$), finally Z_w are concatenated to form the multi-stage fusion $h_{multi-stage}$.

Uni-stage Multimodal Fusion. The uni-stage fusion is applied to learn the different contributions of the multimodal feature at each temporal and produce a summary, which is regarded as the global multimodal fusion. We use another block to learn uni-stage multimodal fusion. Specifically, as shown in Eq. (7), we use a non-linear projection layer to project features of three modalities into the same space.

$$h_i^{'} = f\left(W_f\left[h_i^a, h_i^v, h_i^l\right] + b_f\right) \tag{7}$$

where W_f is the trainable weights, b_i is the bias, f is $relu$ activation function and [ ] denotes for concatenation. Then we perform attention operation on the projected results $h_i^{'}$ to get a summary about of which stages the multimodal features are most important for sentiment analysis, as shown in Eqs. (8–10).

$$o_i = tanh\left(W_a h_i^{'} + b_i\right) \tag{8}$$

$$\alpha_i = \ softmax\left(u_a o_i\right) \tag{9}$$

$$h_{uni-stage} = \sum_{i=1}^{T} a_i h_i^{'} \tag{10}$$

where α_i is the attention weight of timestamp i. We use the attention weights to perform weighted sum on $h_i^{'}$, getting the uni-stage multimodal fusion $h_{uni-stage}$.

3.3 Final Representation and Model Training

As mentioned before, we believe that language modality contains richer information than other modalities, thus we perform attention operation on H_L to get the final language representation h_l. At last we concatenate h_l, the multi-stage multimodal fusion $h_{multi-stage}$ and uni-stage multimodal fusion $h_{uni-stage}$ to form the final representation h_{final}. The final representation is input to a fully-connected layer and a prediction layer to get the output, as shown in Eqs. (11–12):

$$h_{final}^{'} = relu\left(W_1 h_{final} + b_1\right) \tag{11}$$

$$y = f(W_2 h_{final}^{'} + b_2) \tag{12}$$

where W_1 and W_2 are trainable weights, b_1 and b_2 are biases. f is $softmax$ function for classification task. For regression task, we don't need activation function. y is the prediction.

4 Experiments

4.1 Dataset

We conduct several experiments on the CMU-MOSI [26] dataset, the CMU-MOSEI [24] dataset and the YouTube [10] dataset. The CMU-MOSI dataset contains 93 videos from the social media website, each of which comes from a different speaker who is expressing his or her opinions towards a movie. The videos in CMU-MOSI dataset are split into 2199 video clips, and each clip has a sentiment label $y \in [-3, 3]$, which represents strongly positive (labeled as $+3$), positive ($+2$), weakly positive ($+1$), neutral (0), weakly negative (-1), negative (-2), strongly negative (-3) respectively. The CMU-MOSEI dataset is a made up of 23,043 movie review video clips taken from YouTube. Following [8], we consider positive, negative and neutral sentiments in the paper. The YouTube dataset is collected from YouTube which contains 269 video clips. The statistical information of the three datasets is shown in Table 1.

Table 1. The statistical information of the experimental dataset.

Dataset	CMU-MOSI	CMU-MOSEI	YouTube
#Train	1284	15920	173
#Valid	229	2291	36
#Test	686	4832	60

4.2 Evaluation Metrix

Following previous work, we use different evaluation metrix on different datasets. For CMU-MOSI, we conduct experiments on binary classification task, multi-class classification task and regression task. For binary classification, we report accuracy and F1 score, whereas for multi-class classification we only report accuracy. For regression task, we report Mean Absolute Error (MAE) and Pearson's Correlation (Corr). For all the metrics, higher values denote better performance, except MAE where lower values denote better performance. For CMU-MOSEI and YouTube datasets, we conduct 3 classification task and report accuracy and F1 score.

4.3 Experimental Details

For all datasets, 300-dimensional GloVe embeddings [13] are used to represent the language features; Facet[1] library is used to extract a set of visual features

[1] https://imotions.com/biosensor/fea-facial-expression-analysis/.

and COVAREP [3] is used to extract acoustic features. We use WordNet to get 4 sense words for each word. Note that we add a constraint that the sense words should contain the original word. Besides, in WordNet, sense may contains more than one word, if this happen we use the average embedding of the words in the sense as the representation of the sense. The sizes of hidden states of LSTMs encoding language features, vision features and acoustic features are 100, 10 and 30 respectively. We use CNNs with window size 1 and 3 respectively to learn the multi-stage multimodal fusion and the filter number of CNN is set to 50. The batch size is set to 32, 16 and 16 for CMU-MOSI, CMU-MOSEI, YouTube datasets respectively, and the initial learning rate is set to 0.0008, 0.0003 and 0.0001 for the three datasets respectively. For CMU-MOSI dataset, we use L1 loss as training loss, for other two datasets, we use cross entropy loss as training loss. We report the experimental results predicted by the model which performs best on the validation set.

4.4 Baseline Models

We use several models as our baselines to compare with our model. Firstly, we use THMM [10] and MV-HCRF [18] as the traditional baseline models. THMM [10] concatenates language, acoustic and vision features and then uses HMM for classification. MV-HCRF [18] is an extension of the HCRF for Multi-view data, explicitly capturing view-shared and view specific sub-structures. Secondly, we use MV-LSTM [17], BC-LSTM [15], CAT-LSTM [16], GME-LSTM [2], TFN [22], CHFusion [9], LMF [7], MFN citech26ZadehLMPCM18, RMFN [6] and MARN [25] as the early neural network based compared models. Lastly, we use several previous state of the art models as our baseline models. MCTN [14] learns joint representations of multi-modalities by cyclic translations between modalities. HFFN [8] proposes a hierarchical feature fusion network, named 'divide, conquer and combine' to explore both local and global interactions in multiple stages. MFM [19] is proposed to optimize for a joint generative-discriminative objective across multimodal data and labels.

4.5 Experimental Results

Experimental Results on the CMU-MOSI Dataset. The results of our model and baseline models on the CMU-MOSI dataset is shown in Table 2. As is shown, the neural network based models outperform traditional machine learning models with a large margin. Among all models, our model achieves the second best performance on accuracy and F1 score of binary classification and accuracy of 7 classification, and our model achieves the best performance on MAE and Pearson's correlation of regression task compared with the baseline models. Specifically, our model achieves competitive results compared with HFFN on binary classification task, and outperforms MCTN, which is the best model on MAE among the baseline models by 4.5% on MAE. For Pearson's correlation (Corr), our model outperforms RMFN which achieves the best performance on Corr among the baselines by 3.9%. As for seven classification task,

we achieve the second best performance. The overall experimental results on the CMU-MOSI dataset show the effectiveness of our model.

Table 2. Experimental results of different models on the CMU-MOSI dataset.

Model	Binary		Regression		7-class
	Acc	F1	MAE	Corr	Acc
THMM [10]	50.7	45.4	–	–	17.8
MV-HCRF [18]	65.6	65.7	–	–	24.6
MV-LSTM [17]	73.9	74.0	1.019	0.601	33.2
BC-LSTM [15]	73.9	73.9	1.079	0.581	28.7
GME-LSTM [2]	76.5	73.4	0.955	–	–
TFN [22]	74.6	74.5	1.040	0.587	28.7
LMF [7]	76.4	75.7	0.912	0.668	32.8
RMFN[6]	78.4	78.0	0.922	0.681	**38.3**
MARN [25]	77.1	77.0	0.968	0.625	34.7
MFN [23]	77.4	77.3	0.965	0.632	34.1
MFM [19]	78.1	78.1	0.951	0.662	36.2
MCTN [14]	79.3	79.1	0.909	0.676	–
HFFN [8]	**80.2**	**80.3**	–	–	–
MPFN(Ours)	80.0	80.0	**0.864**	**0.720**	37.0

Experimental Results on the YouTube Dataset. Table 3 shows the experimental results of our model and the baseline models on the YouTube dataset. The YouTube is a very small dataset, as shown in Table 1, not all neural network based models outperform traditional machine learning models both on accuracy and F1 score. However, compared with the baseline models, our model achieves the best performance on both accuracy and F1 score, which outperforms the previous state-of-the-art model MFM by 1.7% on accuracy and 3.5% on F1 score. Although the YouTube dataset is very small, our model can achieve the best performance among the baseline models.

Experimental Results on the CMU-MOSEI Dataset. For the CMU-MOSEI dataset, we conduct experiments on 3 classification tasks. We present the experimental results of different models in Table 4. As we can see, our model achieves the best performance on both accuracy and F1 score, which outperforms HFFN by 0.93% on accuracy and 0.6% on F1 score, and outperforms BC-LSTM by 0.53% on accuracy and 0.63% on F1 score. Note that the CMU-MOSEI is the largest dataset in this paper. In addition, we can see that although CAT-LSTM and LMF achieve relative good performance on accuracy, their performance on F1 score is much worse than that on accuracy. Our model can achieve both good performance on accuracy and F1 score. Experimental results on the CMU-MOSEI dataset and the YouTube dataset show that our model can adapt to both small data and large data.

Table 3. Experimental results of different models on the YouTube dataset

Model	Acc	F1
THMM [10]	42.4	27.9
MV-HCRF [18]	44.1	44.0
MV-LSTM [17]	45.8	43.3
BC-LSTM [15]	45.0	45.1
TFN[22]	45.0	41.0
MARN [25]	48.3	44.9
MFN [23]	51.7	51,6
MCTN [14]	51.7	52.4
MFM [19]	53.3	52.4
MPFN(Ours)	**55.0**	**55.9**

Table 4. Experimental results of different models on the CMU-MOSEI dataset.

Model	Acc	F1
BC-LSTM [15]	60.77	59.04
TFN [22]	59.40	57.33
CAT-LSTM [16]	60.72	58.83
CHFusion [9]	58.45	56.90
LMF [7]	60.27	53.87
HFFN [8]	60.37	59.07
MPFN(Ours)	**61.30**	**59.67**

4.6 Ablation Studies

In order to investigate the impact of various components in our model, we conduct several ablation experiments on the CMU-MOSI dataset, which are shown in Table 5. In the experiment, we remove one kind of component of our full model each time. Specifically, we remove the sense-level attention (denoted as MPFN-no-sense-att), the multi-stage multimodal fusion (denoted as MPFN-no-multi-stage-fusion), the uni-stage multimodal fusion (denoted as MPFN-no-uni-stage-fusion) and final language representation (denoted as MPFN-no-language-final) respectively.

As shown in Table 5, once we remove any component of our model, the performance will decline. For example, if we remove the sense-level attention and use the original word embedding as word representation, the performance of our model will drop by 1.0% on accuracy, 1.4% on F1 score of binary classification task, 3.8% on MAE, 2.5% on Corr, and 2.9% on accuracy of 7 classification task on the CMU-MOSI dataset. This observation suggests that using WordNet and sense-level attention to dynamically learn the word representation is effective.

In terms of multimodal fusion, we can see that if we remove the multi-stage fusion block or the uni-stage fusion block, the performance of our model will also drop, which indicates that both multi-stage fusion and uni-stage fusion are important for multimodal sentiment analysis. Furthermore, it seems that the multi-stage multimodal fusion plays a more important role than uni-stage multimodal fusion on the CMU-MOSI dataset.

Last but not least, we remove the final language representation which is concatenated with the multimodal fusion representation to see whether this operation is useful. The experimental results prove our early assumption. As we mentioned, ablation studies of previous researches show that if only using features of one modality as input, the model which use language modality features as input performs best. If only using multimodal fusion representation to form the final representation, some intra-modality information of language will be lost during fusion process. Concatenating the final language representation with the multimodal fusion representation to form the final representation can address this problem.

Table 5. Ablation studies on the CMU-MOSI dataset.

Model	Binary		Regression		7-class
	Acc	F1	MAE	Corr	Acc
MPFN-no-sense-att	79.0	78.6	0.902	0.695	34.1
MPFN-no-multi-stage-fusion	79.0	79.0	0.882	0.698	36.9
MPFN-no-uni-stage-fusion	79.3	79.3	0.888	0.711	33.5
MPFN-no-language-final	79.3	79.3	0.899	0.714	34.4
MPFN(Ours)	**80.0**	**80.0**	**0.864**	**0.720**	**37.0**

4.7 Discussion

In order to investigate how each modality effects the performance of our model, we conduct several experiments to compare the performance of our model using unimodal, bimodal and multimodal features, as shown in Table 6.

For unimodal features, we can see that our model only using sense attentive language representation outperforms the model that only using audio features or video features with significant margin, which is consistent with our early assumption that language modality is dominant. For bimodal features, we can infer that when integrating language modality with acoustic modality or vision modality, the performance of the model outperforms that of only using language representation, which indicates that acoustic and vision modalities play auxiliary roles and the multi-perspective multimodal fusion can improve the performance of the model. However, when using audio features and video features as input, the performance of the model is still much worse than that of only using language

Table 6. The performance of our model using unimodal, bimodal and multimodal features.

Modality	Source	Binary		Regression		7-class
		Acc	F1	MAE	Corr	Acc
Unimodal	Audio	57.1	56.2	1.396	0.196	16.0
	Video	57.3	57.3	1.431	0.137	16.2
	Sense attentive language	79.0	79.1	0.922	0.689	34.0
Bimodal	Sense attentive language + Audio	79.7	79.6	0.881	0.701	34.7
	Sense attentive language + Video	79.6	79.6	0.915	0.714	32.9
	Audio + Video	59.0	59.0	1.391	0.176	19.7
Multimodal	**Sense attentive language + Audio + Video**	**80.0**	**80.0**	**0.864**	**0.720**	**37.0**

modality, which again proves that language modality is the most important modality in this task.

When cooperating three modalities, our full model MPFN achieves the best performance among the different combinations, which demonstrates the effectiveness of multi-perspective multimodal fusion proposed in this paper.

5 Conclusion

In this paper, we propose a novel multi-perspective fusion network focusing on sense attentive language for multimodal sentiment analysis. Evaluations show that using our proposed multi-stage and uni-stage fusion strategies and using sense attentive language representation can improve performance on multimodal sentiment analysis for the CMU-MOSI, CMU-MOSEI and YouTube data. Our model also achieves a new state-of-the-art in the YouTube and CMU-MOSEI dataset on accuracy and F1 measure metrics compared with the baseline models. The experimental results using different modal combinations also show that the proposed sense attentive language modal achieves the most significant performance improvement on the CMU-MOSI dataset, especially on the 7-classification results, indicating that the sense attentive language modal plays an important role in multimodal sentiment analysis task. Like most of other models, our approach also focuses on the multimodal data with the same length of stamp. In the future, we will investigate a novel fusion of multimodal data with different length of stamp.

Acknowledgments. This work is supported by National Natural Science Foundation of China (No. 61976062) and the Science and Technology Program of Guangzhou (No. 201904010303).

References

1. Baltrusaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2019)

2. Chen, M., Wang, S., Liang, P.P., Baltrusaitis, T., Zadeh, A., Morency, L.: Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, pp. 163–171 (2017)
3. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP - a collaborative voice analysis repository for speech technologies. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, pp. 960–964 (2014)
4. Dong, Z.: Knowledge description: what, how and who. In: Proceedings of International Symposium on Electronic Dictionary, vol. 18 (1988)
5. Lazaridou, A., Pham, N.T., Baroni, M.: Combining language and vision with a multimodal skip-gram model. In: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015, pp. 153–163 (2015)
6. Liang, P.P., Liu, Z., Zadeh, A., Morency, L.: Multimodal language analysis with recurrent multistage fusion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 150–161 (2018)
7. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.: Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, pp. 2247–2256 (2018)
8. Mai, S., Hu, H., Xing, S.: Divide, conquer and combine: hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pp. 481–492 (2019)
9. Majumder, N., Hazarika, D., Gelbukh, A.F., Cambria, E., Poria, S.: Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowl. Based Syst. **161**, 124–133 (2018)
10. Morency, L., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: harvesting opinions from the web. In: Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, pp. 169–176 (2011)
11. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, pp. 689–696 (2011)
12. Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic inquiry and word count: Liwc [computer software]. Austin, TX: liwc. net 135 (2007)
13. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, pp. 1532–1543 (2014)
14. Pham, H., Liang, P.P., Manzini, T., Morency, L., Póczos, B.: Found in translation: learning robust joint representations by cyclic translations between modalities. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, pp. 6892–6899 (2019)
15. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.: Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 873–883 (2017)
16. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.: Multi-level multiple attentions for contextual multimodal sentiment analysis. In: 2017 IEEE International Conference on Data Mining, ICDM 2017, pp. 1033–1038 (2017)

17. Rajagopalan, S.S., Morency, L.-P., Baltrušaitis, T., Goecke, R.: Extending long short-term memory for multi-view structured learning. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 338–353. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_21
18. Song, Y., Morency, L., Davis, R.: Multi-view latent variable discriminative models for action recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2120–2127 (2012)
19. Tsai, Y.H., Liang, P.P., Zadeh, A., Morency, L., Salakhutdinov, R.: Learning factorized multimodal representations. In: 7th International Conference on Learning Representations, ICLR 2019 (2019)
20. Wang, Y., Shen, Y., Liu, Z., Liang, P.P., Zadeh, A., Morency, L.: Words can shift: dynamically adjusting word representations using nonverbal behaviors. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, pp. 7216–7223 (2019)
21. Xie, R., Yuan, X., Liu, Z., Sun, M.: Lexical sememe prediction via word embeddings and matrix factorization. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 4200–4206 (2017)
22. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, pp. 1103–1114 (2017)
23. Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.: Memory fusion network for multi-view sequential learning. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 5634–5641 (2018)
24. Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.: 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, pp. 2236–2246 (2018)
25. Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E., Morency, L.: Multi-attention recurrent network for human communication comprehension. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 5642–5649 (2018)
26. Zadeh, A., Zellers, R., Pincus, E., Morency, L.: MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. Computing Research Repository arXiv:1606.06259 (2016)
27. Zeng, X., Yang, C., Tu, C., Liu, Z., Sun, M.: Chinese LIWC lexicon expansion via hierarchical classification of word embeddings with sememe attention. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 5650–5657 (2018)