

# SWAFN: Sentimental Words Aware Fusion Network for Multimodal Sentiment Analysis

Minping Chen, Xia Li\*

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

## Abstract

Multimodal sentiment analysis aims to predict sentiment of language text with the help of other modalities, such as vision and acoustic features. Previous studies focused on learning the joint representation of multiple modalities, ignoring some useful knowledge contained in language modal. In this paper, we try to incorporate sentimental words knowledge into the fusion network to guide the learning of joint representation of multimodal features. Our method consists of two components: shallow fusion part and aggregation part. For the shallow fusion part, we use crossmodal coattention mechanism to obtain bidirectional context information of each two modals to get the fused shallow representations. For the aggregation part, we design a multitask of sentimental words classification to help and guide the deep fusion of the three modalities and obtain the final sentimental words aware fusion representation. We carry out several experiments on CMU-MOSI, CMU-MOSEI and YouTube datasets. The experimental results show that introducing sentimental words prediction as a multitask can really improve the fusion representation of multiple modalities.

## 1 Introduction

Multimodal sentiment analysis is a task of predicting the sentiment of a video, an image or a text based on multiple modal features. Based on the contributions of different modalities to each other, multimodal sentiment analysis has achieved significant results and attracted the attentions of many researchers in recent years.

The main challenge of the multimodal sentiment analysis is to capture a better fusion of different modalities. Previous studies have proposed different methods for the fusion in different point of views. Some methods focus on the improvement of the LSTM structure to learn the interactions of different modal features from the view of the uni-stage and multi-stage. Zadeh et al. (2018a) propose a Memory Fusion Network to learn both the view-specific interactions and the cross-view interactions. Liang et al. (2018) propose a Recurrent Multistage Fusion Network to model cross-modal interactions using multi-stage fusion approach. Some methods focus on exploiting the expressiveness of tensors for multimodal representation. Zadeh et al. (2017) propose a Tensor Fusion Network to explicitly model the unimodal, bimodal and trimodal interactions through a 3-fold Cartesian product from modality embedding. More recently, other methods are proposed and achieve new state-of-the-art results (Pham et al., 2019; Mai et al., 2019; Wang et al., 2019; Tsai et al., 2019).

Although previous studies achieved good results, there are still two points can be improved: (1) We find that the fusion of most of previous methods is from one direction, that is, when the two modals are fused, the representation of two modals are fused directly as a new representation, similar to the work of Zadeh et al. (2017) and Liu et al. (2018). This fusion strategy ignores the long range of context information of each modality. As an example shown in Figure 1, for the fusion of language modality and vision modality, if we can capture the context information of each modality from bi-directions, we can

\*Corresponding author: xiali@gdufs.edu.cn

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.



Figure 1: Contexts of each modality captured by crossmodal coattention. (a) shows each language temporal captures context of vision modality by different attention weights. (b) shows each vision temporal captures context of language modality by different attention weights.

get more sufficient fusion information. (2) Few of previous studies explicitly explored the knowledge contained in the language text which can be used to help the fusion of different modalities based on the rich information existing in the language.

To this end, in this paper, we propose a Sentimental Words Aware Fusion Network (SWAFN) for multimodal sentiment analysis. More specifically, we first use LSTM to encode the original features of three modalities. Then we use the coattention mechanism (Xiong et al., 2017) to learn the co-dependent representation between language and other modalities separately by capturing attention contexts of each modality. We call this kind of bimodal fusion between language and other modalities as the shallow fusion part. Figure 1 presents the illustration of crossmodal coattention for language and vision modalities. Then, we design a sentimental words prediction task as an auxiliary task through the multitask learning mechanism to guide the aggregation of the shallow fusion of multiple modal features and obtain the final sentimental words aware deep fusion representation.

The main contribution of this work are as follows:

1) We propose to use crossmodal coattention to learn the long range context information of each two modals to obtain more sufficient fusion information for multiple modals. We also design a sentimental words prediction multitask as an auxiliary task to guide the fusion of multiple modal features and learn sentimental words aware final representation. To the best of our knowledge, this is the first time that multi-task learning is applied in multimodal sentiment analysis.

2) We conduct several experiments on different public datasets, and we will show that our model is effective for multimodal sentiment analysis. In addition, we also carry out a series of experiments to investigate the contribution of different modalities, the impact of the shallow fusion and the final fusion after integrating the auxiliary task.

## 2 Related Work

The key problem of multimodal sentiment analysis is to fill the gap of different modalities and learn the effective fusion of multimodal features. In recent years, with the successful application of neural networks in many tasks, different sophisticated fusion approaches are proposed and achieve significant results.

**Fusion methods based on improved LSTM structure.** Some of the previous studies propose to improve the LSTM structure to learn the interactions of different modality features from the view of the same timestep and cross timestep. Chen et al. (2017) propose a Gated Multimodal Embedding LSTM with Temporal Attention model which consists of two modules, one is Gated Multimodal Embedding aiming to alleviate the fusion difficulty when there are noisy modalities, another is LSTM with temporal attention to perform word-level fusion. Zadeh et al. (2018c) propose a Multi-attention Recurrent Network, in which the LSTHM (an extension of LSTM) is used to store view-specific dynamics of the assigned modality and cross-view dynamics related to the assigned modality, and the Multi-attention Block is used to discover cross-view dynamics cross different modalities. Zadeh et al. (2018a) propose a Memory Fusion Network which employs LSTM to learn view-specific interactions and an attention mechanism called the Delta-memory Attention Network to identify the cross-view interactions. Liang et al. (2018) propose a Recurrent Multistage Fusion Network to model cross-modal interactions using multi-stage fusion approach.

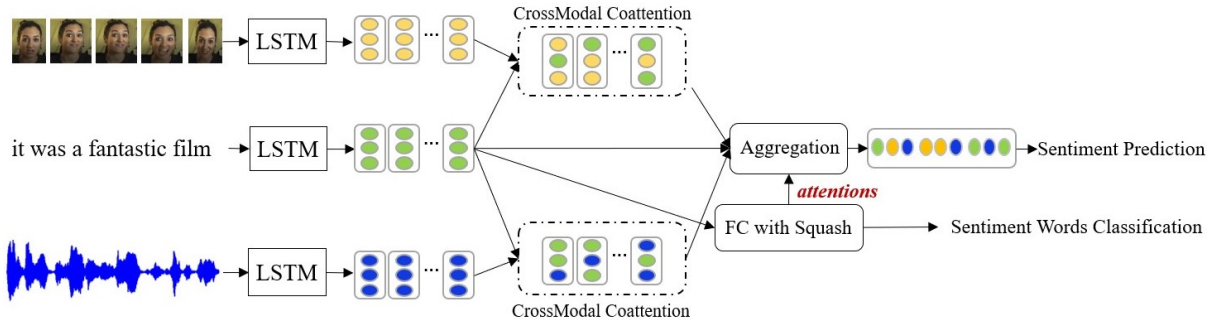


Figure 2: The whole architecture of our model. We first use coattention mechanism to learn the bidirectional long range of context information between language modality and other modalities separately. Then we integrate a sentimental words classification task into the model through multitask learning mechanism to guide the learning and aggregation of multimodal fusion.

**Fusion methods based on tensor structure.** Different from using the improved LSTM-based models, some previous studies exploit the expressiveness of tensors for multimodal representation. Zadeh et al. (2017) propose a Tensor Fusion Network to explicitly model the interactions of different modals through a 3-fold Cartesian product from modality embedding. Liu et al. (2018) propose a Low-rank Multimodal Fusion network which first obtain the unimodal representation and perform low-rank multimodal fusion to improve the efficiency.

**Previous state-of-the-art fusion methods.** More recently, Pham et al. (2019) explore a method of translations between modalities to learn joint representations, in which a cycle consistency loss is used to ensure that the joint representations retain maximal information from all modalities. Different from most previous studies which directly fuse features at holistic level, Mai et al. (2019) propose a “divide, conquer and combine” strategy to perform multimodal fusion hierarchically which considers both local and global interactions. In order to model expressive nonverbal representations, Wang et al. (2019) propose a Recurrent Attended Variation Embedding Network which analyzes the fine-grained visual and acoustic patterns and dynamically shifts word representations according to nonverbal cues. Tsai et al. (2019) introduce a Multimodal Factorization Model which factorizes representations into two sets of independent factors: multimodal discriminative and modality-specific generative factors and propose a joint generative-discriminative objective to optimize across multimodal data and labels.

Although previous studies have proposed many effective multimodal fusion approaches, few studies have explored the possibility of using knowledge in language as a multi-task learning framework in multimodal sentiment analysis. In this paper, we try to design an auxiliary task to guide the model to learn sentimental words information aware multimodal representation.

### 3 Our Model

In this section, we will describe our model in more detail. Section 3.1 introduces the crossmodal coattention, section 3.2 introduces the sentimental words prediction auxiliary task, section 3.3 introduces the sentimental words aware representation and section 3.4 describes the model training. Figure 2 shows the whole architecture of our model.

#### 3.1 CrossModal Coattention

Given the word embedding of language, the raw features of acoustic and vision modalities, denoted as  $X_L = \{l_1, l_2, \dots, l_T\}$ ,  $X_A = \{a_1, a_2, \dots, a_T\}$  and  $X_V = \{v_1, v_2, \dots, v_T\}$  respectively, we use LSTM to model the temporal information of the three modalities as intra-modal encoding, getting the LSTM hidden states output of the three modalities, denoted as  $H_L$ ,  $H_A$  and  $H_V$  respectively.

After getting the encoded features of three modalities  $H_L$ ,  $H_A$  and  $H_V$  respectively, we use coattention (Xiong et al., 2017) to learn the bimodal fusion between language modality and other modalities. Firstly, we use a non-linear projection layer to transform the dimension of the encoded language repre-

sentation into the same dimension of that of other modalities in order to perform coattention, as show in equation(1).

$$H'_L = \tanh(H_L W_L + b_L) \quad (1)$$

The coattention mechanism is applied to attend to the language modality and other modality(i.e. vision or acoustics) simultaneously, and learn the bimodal fusion. Firstly, an affinity matrix is computed, which contains the affinity scores corresponding to all pairs of language hidden states and vision(or acoustic) hidden states. Then the softmax function is used to normalize the affinity matrix row-wise to produce the attention weights  $A_V$ (or  $A_L$ ) across the language text for each timestep of the vision(or acoustic) features, and column-wise to produce the attention weights  $A_L$  across the vision(or acoustic) features for each word, as shown in equation(2-4):

$$\alpha = H_V (H'_L)^T \quad (2)$$

$$A_V = \text{softmax}(\alpha) \quad (3)$$

$$A_L = \text{softmax}(\alpha^T) \quad (4)$$

Next, we compute the attention contexts of the language features based on the attention weights of each timestep of the vision(or acoustic) features, as shown in equation(5):

$$C_V = A_V H'_L \quad (5)$$

Similarly, we can compute the attention contexts  $A_L H_V$  of the vision(or acoustic) features based on the attention weights of each word of the language features. Following the work of (Xiong et al., 2017), we also compute the summaries  $A_L C_V$  to map the vision(or acoustic) features encoding into the space of language features encoding. The corresponding operation is shown in equation(6):

$$C_{L\&V} = A_L [H_V, C_V] \quad (6)$$

Where  $C_{L\&V}$  is defined as a co-dependent representation of the language modality and vision modality.  $[ ]$  denotes for concatenation operation. Similarly, we can get  $C_{L\&A}$  using the same coattention operation for language modality and acoustic modality as a co-dependent representation of the language modality and acoustic modality. The bimodal fusion  $C_{L\&V}$  and  $C_{L\&A}$  are regarded as a kind of shallow fusion, as the trimodal fusion and the knowledge existing in the language modality are not well captured so far.

### 3.2 Sentimental Words Prediction Auxiliary Task

In addition to use other modalities to assist language modality, we find that the sentimental words information existing in the language modal can also be incorporated into the fusion model to learn richer multimodal representation. In this paper, we design a word-level classification task which is used to determine whether each word is a sentimental word. Specifically, we use Bing Liu’s Opinion Lexicon as the knowledge<sup>1</sup>, which contains the negative-words list and positive-words list to obtain the label of each word. We first merge the two lists into a sentimental word list. If a word is in the sentimental word list, then it is a sentimental word, otherwise, it is not a sentimental word. Then we build the auxiliary task as a multi-label classification task as each sentence in the language modality may contain more than one sentimental word. Note that the word-level classification task and the sentiment analysis task share the same language encoding layer, as shown in Figure 2. We input  $H_L$  into a fully-connected layer with a row-wise squash activation function (Sabour et al., 2017) to adjust it to prepare for word-level classification. The squash function is used to ensure that short vectors get shrunk to almost zero length and long vectors get shrunk to a length slightly below 1. We expect that the squash function can learn representation where the length of the vector can represent the probability of each word to be a sentimental word.

<sup>1</sup><http://sentiment.christopherpotts.net/lexicons.html>

The operation is shown in (7-8).

$$\text{squash}(x) = \frac{\|x\|^2}{1+\|x\|^2} \frac{x}{\|x\|} \quad (7)$$

$$H_{words} = \text{squash}(H_L W_w + b_w) \quad (8)$$

Where  $W_w$  is weight and  $b_w$  is bias. Then  $H_{words}$  is input to a multi-label classification layer, as shown in equation (9).

$$y_{words} = \text{softmax}(H_{words} W_{words} + b_{words}) \quad (9)$$

Where  $W_{words}$  is weight and  $b_{words}$  is bias.  $y_{words} \in \mathcal{R}^T$ , which denotes whether each word is sentimental word.

### 3.3 Sentimental Words Aware Multimodal Representation

As described in section 3.1, we get the bimodal fusion between language with other modalities separately. As mentioned earlier, we view this fusion as a shallow fusion because we believe that there is rich semantic information in the language which can be fused to learn the deep fusion and aggregation of different modalities. As demonstrated by many previous studies (Poria et al., 2017a; Zadeh et al., 2017; Mai et al., 2019), the language modality often plays a dominated role among the three modalities, thus we concatenate  $C_{L\&V}$ ,  $C_{L\&A}$  and  $H_L$ , and input the result to a LSTM layer to aggregate the two kinds of bimodal fusion representation and the intra-modality encoding of language, getting  $H_{agg}$ , as shown in equation (10).

During the training of the auxiliary task, we expect that the  $H_{words} = \{h_1^w, h_2^w, \dots, h_T^w\}$  can learn the information about whether each word is sentimental word and the representation of the sentimental words can be distinguished from that of non-sentimental words. For sentiment analysis, the sentimental words are usually the key clues for determining the sentiment. However, in some cases, a sentence may contain sentimental words with different polarities and we need to decide which sentimental words make more contribution for sentiment prediction. Thus, to enable the auxiliary task to produce a marked effect, we use the final representation of word-level representation  $H_{words}$  to learn the contribution of each word and guide the learning of multimodal fusion, as shown in equations (11-13).

$$H_{agg} = \text{LSTM}([C_{L\&V}, C_{L\&A}, H_L]) \quad (10)$$

$$o_i = \tanh(h_i^w W_a + b_a) \quad (11)$$

$$\alpha_i = \text{softmax}(o_i W_u) \quad (12)$$

$$S_{att} = \sum_{i=1}^T \alpha_i h_i^{agg} \quad (13)$$

Where  $W_a$  and  $W_u$  are trainable weights and  $b_a$  is the bias. Note that we use the learned attention weights to perform weighted sum on the multimodal fusion representation  $H_{agg}$ , getting  $S_{att}$ , which is the sentimental words information aware representation.

In addition to  $S_{att}$  learned through the guiding of the auxiliary task, we perform average pooling on  $H_{agg}$  to obtain the global multimodal information, which is denoted as  $S_{avg}$ . Finally, we concatenate  $S_{att}$  and  $S_{avg}$  to form the final representation. The final representation is input to a fully-connected layer and a prediction layer to get the sentiment prediction, as shown in equations (14-15).

$$S = [S_{att}, S_{avg}] W_f + b_f \quad (14)$$

$$y_s = S W_s + b_s \quad (15)$$

### 3.4 Model Training

Considering that classification task and regression task for sentiment analysis are simultaneously evaluated on CMU-MOSI dataset, we use L1 Loss for training the sentiment analysis tasks of CMU-MOSI

dataset, which is shown in equation (16). Where  $y_s^i$  and  $\hat{y}_s^i$  are the true sentiment and predicted sentiment of  $i$ -th sample respectively.  $N$  is number of training samples. For CMU-MOSEI and YouTube datasets, following (Pham et al., 2019), we use Cross-entropy as training loss.

$$Loss_{sa} = \frac{1}{N} \sum_{i=1}^N |y_s^i - \hat{y}_s^i| \quad (16)$$

As for the word-level classification task, we use Binary Cross Entropy Loss, which is shown in equation (17). Where  $y_w^i$  and  $\hat{y}_w^i$  are the true label and the predicted label of the  $i$ -th sample respectively,  $T$  is the length of language sentence.

$$Loss_{sw} = \frac{1}{N} \sum_{i=1}^N \left\{ -\frac{1}{T} \sum_{j=1}^T (y_w^{ij} * \log(\hat{y}_w^{ij}) + (1 - y_w^{ij}) * \log(1 - \hat{y}_w^{ij})) \right\} \quad (17)$$

The overall loss of our model is the weighted sum of  $Loss_{sa}$  and  $Loss_{sw}$ , as shown in equation (18), where  $\alpha \in (0, 1)$  is a hyper parameter.

$$Loss = (1 - \alpha) * Loss_{sa} + \alpha * Loss_{sw} \quad (18)$$

## 4 Experiments

### 4.1 Dataset

We use CMU-MOSI, CMU-MOSEI and YouTube as our experimental datasets, which are extensively used in the previous studies. Following most previous studies, GloVe embeddings (Pennington et al., 2014) are used to represent the language features, the visual features are extracted by Facet library<sup>2</sup> and acoustic features are extracted using COVAREP (Degottex et al., 2014).

CMU-MOSI (Zadeh et al., 2016) contains 93 videos from YouTube, each of the videos is expressing a speaker’s opinions towards a movie. The videos are split into 2199 clips. We train our model on 52 videos (1284 clips), validates on 10 videos (229 clips) and tests on 31 videos (686 clips). Each sentiment label of the clip is a number between  $[-3, 3]$ , which represents strongly positive (denoted as +3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2), strongly negative (-3) respectively. CMU-MOSEI (Zadeh et al., 2018b) consists of 22,413 video clips about movie reviews from YouTube. There are 15290, 2291 and 4832 clips in the training set, validation set and test set respectively. YouTube (Morency et al., 2011) consists of 269 video clips, in which the size of training set, validation set and test set are 173, 36 and 60 respectively.

For CMU-MOSI dataset, we complete binary classification, multi-class classification and regression experiments. For regression task, we report Mean Absolute Error (MAE) and Pearson’s Correlation (Correlation). For binary classification, we report accuracy and F1 score, while for multi-class classification we only report accuracy, which is consistent with most previous studies. For CMU-MOSEI and YouTube dataset, we consider positive, negative and neutral sentiments following (Mai et al., 2019) and use accuracy and F1 score. For all metrics, higher values represent better performance, except for MAE.

### 4.2 Settings

We use 300-dimensional GloVe (Pennington et al., 2014) word embeddings as language features. The hidden sizes of LSTMs encoding of language, vision and acoustic features for CMU-MOSI dataset are 100, 30 and 50 respectively. The same hidden sizes for CMU-MOSEI dataset are 128, 10 and 20 respectively, for YouTube dataset are 100, 20 and 20 respectively. The batch size is set to 16, 64 and 16 for CMU-MOSI, CMU-MOSEI and YouTube datasets respectively. We set hidden sizes of LSTM for aggregating the multimodal fusion to 128, 100 and 100, the trade-off parameter  $\alpha$  between the sentiment prediction loss and word-level classification loss to 0.3, 0.25 and 0.25, the initial learning rate to  $6e-4$ ,  $4e-4$  and  $5e-4$  for the three datasets respectively. The hidden sizes of the fully-connected layer before

<sup>2</sup><https://imotions.com/biosensor/fea-facial-expression-analysis/>

Model	Binary		Regression		7-class
	Acc	F1	MAE	Corr	Acc
MV-LSTM(Rajagopalan et al., 2016)	73.9	74.0	1.019	0.601	33.2
BC-LSTM (Poria et al., 2017a)	73.9	73.9	1.079	0.581	28.7
GME-LSTM (Chen et al., 2017)	76.5	73.4	0.955	-	-
TFN (Zadeh et al., 2017)	74.6	74.5	1.040	0.587	28.7
LMF (Liu et al., 2018)	76.4	75.7	0.912	0.668	32.8
RMFN (Liang et al., 2018)	78.4	78.0	0.922	0.681	38.3
MARN (Zadeh et al., 2018c)	77.1	77.0	0.968	0.625	34.7
MFN (Zadeh et al., 2018a)	77.4	77.3	0.965	0.632	34.1
MFM (Tsai et al., 2019))	78.1	78.1	0.951	0.662	36.2
MCTN (Pham et al., 2019)	79.3	79.1	0.909	0.676	-
HFFN (Mai et al., 2019)	<b>80.2</b>	<b>80.3</b>	-	-	-
SWAFN(Ours)	<b>80.2</b>	80.1	<b>0.880</b>	<b>0.697</b>	<b>40.1</b>

Table 1: Experimental results of different models on CMU-MOSI dataset.

the prediction layer are set to 100, 100 and 200 on CMU-MOSI, CMU-MOSEI, and YouTube datasets respectively. The proposed model is trained for 20 epoch, 8 epoch and 25 epoch on the three datasets respectively. We select the model which performs best on the validation set to evaluate on the test set. <sup>3</sup>

### 4.3 Baseline Models

We use the following methods as our baseline models for experiments. Firstly, we use MV-LSTM (Rajagopalan et al., 2016), BC-LSTM (Poria et al., 2017a), CAT-LSTM (Poria et al., 2017b), GME-LSTM (Chen et al., 2017), TFN (Zadeh et al., 2017), CHFusion (Majumder et al., 2018), LMF (Liu et al., 2018), MFN (Zadeh et al., 2018a), RMFN (Liang et al., 2018) and MARN (Zadeh et al., 2018c) as our baseline models based on neural networks which are introduced in section 2. Secondly, we use previous state-of-the-art models as our compared models, such as MCTN (Pham et al., 2019), HFFN (Mai et al., 2019) and MFM (Tsai et al., 2019).

### 4.4 Experimental Results

In this section we present the experimental results and the analysis of our model on CMU-MOSI, YouTube and CMU-MOSEI datasets.

**Experimental results on CMU-MOSI dataset.** We summarize the experimental results of different models on the CMU-MOSI dataset in Table 1. As shown in Table 1, our model achieves competitive performance compared with the best baseline model HFFN on accuracy and F1 score of binary classification. For regression task, our model achieves best performance among the baselines both on mean absolute error(MAE) and correlation(Corr). Specifically, our model outperforms MCTN by 2.9% on MAE and 2.1% on correlation, which are significant improvements. For 7 classification task, our model also achieve the best performance among the baseline models, which outperforms RMFN by 1.8% and MFM by 3.9% on accuracy. The experimental results on CMU-MOSI dataset show that our approach brings more significant improvements on regression task and 7 classification than binary classification task.

**Experimental results on YouTube dataset.** Table 2 shows the experimental results of our model and the baseline models on YouTube dataset. Although the size of YouTube dataset is very small, we can see that compared with the baseline models, our model achieves the best performance on both accuracy and F1 score, which outperforms MCTN by 3.3% on accuracy and 0.9 % on F1 score, and outperforms the previous state-of-the-art model MFM by 1.7% on accuracy and 0.9% on F1 score. Due to the very limited training samples, many baseline models may be overfitting on the training set. Our model achieves better performance, indicating its better generalization ability.

<sup>3</sup>Our source code is released at <https://github.com/gdufslp/SWAFN>

Model	YouTube-Acc	YouTube-F1	MOSEI-Acc	MOSEI-F1
MV-LSTM (Rajagopalan et al., 2016)	45.8	43.3	-	-
BC-LSTM (Poria et al., 2017a)	45.0	45.1	60.77	59.04
TFN (Zadeh et al., 2017)	45.0	41.0	59.40	57.33
CAT-LSTM (Poria et al., 2017b)	-	-	60.72	58.83
MARN (Zadeh et al., 2018c)	48.3	44.9	-	-
MFN(Zadeh et al., 2018a)	51.7	51.6	-	-
CHFusion(Majumder et al., 2018)	-	-	58.45	56.90
LMF(Liu et al., 2018)	-	-	60.27	53.87
MCTN (Pham et al., 2019)	51.7	52.4	-	-
MFN (Tsai et al., 2019)	53.3	52.4	-	-
HFFN (Mai et al., 2019)	-	-	60.37	59.07
<b>SWAFN (Ours)</b>	<b>55.0</b>	<b>53.3</b>	<b>61.03</b>	<b>59.32</b>

Table 2: Experimental results of different models on YouTube dataset and CMU-MOSEI dataset

Modality	Source	Binary		Regression		7-class
		Acc	F1	MAE	Corr	Acc
<b>Unimodal</b>	Audio	57.6	56.7	1.396	0.189	15.3
	Video	58.0	58.1	1.422	0.134	16.2
	language(no auxiliary work)	77.8	77.9	0.931	0.695	35.7
	language(+ auxiliary work)	78.9	78.6	0.903	0.683	36.2
<b>Bimodal</b>	Audio+Video	58.0	58.1	1.384	0.207	15.7
	language +Audio(no auxiliary work)	78.6	78.4	0.906	0.684	35.1
	language +Audio(+ auxiliary task)	79.2	79.2	0.917	0.683	35.1
	language +Video(no auxiliary work)	77.8	77.6	0.921	0.676	35.3
	language +Video(+ auxiliary task)	79.0	78.9	0.882	0.693	37.3
<b>Multimodal</b>	language +Audio+Video(no coattention)	78.4	78.4	0.903	0.688	37.8
	language +Audio+Video(no auxiliary work)	79.2	79.1	0.937	0.682	36.6
	<b>language +Audio+Video(+ auxiliary task)</b>	<b>80.2</b>	<b>80.1</b>	<b>0.880</b>	<b>0.697</b>	<b>40.1</b>

Table 3: The performance of our model using unimodal, bimodal and multimodal features.

**Experimental results on CMU-MOSEI dataset.** For CMU-MOSEI dataset, following (Mai et al., 2019), we conduct experiments on 3 classification tasks. We present the experimental results of different models in Table 2. Our model achieves the best performance on both accuracy and F1 score, which outperforms HFFN by 0.66% on accuracy and 0.25% on F1 score. CMU-MOSEI is the largest dataset among the three datasets, we can see that the difference of the performance of different models is not very significant. For example, the range of the performance on accuracy of the baselines is between 60.2% and 60.8%, except for TFN and CHFusion. However, the range of F1 score of the baseline models is between 53.87% and 59.07% as some baseline models achieve much lower values on F1 score than that on accuracy. Our model can achieve good performance on both accuracy and F1 score. The overall experimental results on three datasets show the effectiveness of our model.

## 5 Discussion

In this section, we investigate the impact of different modalities on the performance of the final model. We also conduct a case study to investigate how the auxiliary task guide the learning of attention weights of sentimental words in the sentence.

### 5.1 Investigation of the Contribution of Different Modalities

In order to investigate the impact of different modalities of our model, we carry out a series of experiments to compare the performance of our model using unimodal, bimodal and multimodal features






















	N	N	N	Y(Neg)	Y(Neg)	N	N	N	Y(Pos)	N	
	And	i	was	unbelievably	shocked	how	much	i	loved	it	Ture: Pos
SWAFN	0.00001	0.00001	0.00002	0.360	0.053	0.00005	0.0006	0.00002	0.585	0.00003	Predicted: Pos
SWAFN( $\Delta$ )	0.00008	0.0001	0.00004	0.000008	0.764	0.132	0.067	0.034	0.0002	0.0001	Predicted: Neg
											
		N	N	N	N	Y(Pos)	Y(Pos)	Y(Neg)			
	Just	give	me	a	nice	interesting	choke		Ture: Pos		
SWAFN	0.00001	0.00005	0.00002	0.00002	0.577	0.207	0.215		Predicted: Pos		
SWAFN( $\Delta$ )	0.004	0.002	0.003	0.004	0.0002	0.0001	0.933		Predicted: Neg		
											

Figure 3: Attention weights learned by SWAFN(our model) and SWAFN( $\Delta$ ) (our model without the sentimental words classification auxiliary task) on two instances. Darker colors indicate greater weights. The auxiliary task can assist the model to pay more attention on the sentimental words and recognize which sentimental words reflect the sentiment correctly.

respectively. We shown them in Table 3.

Firstly, we conduct experiments with the model just using unimodal features, where language(no auxiliary work) is only using language representation and language(+auxiliary work) is fused with sentimental words classifications task. We can see that the model using language modality outperforms the model using acoustic modality or vision modality with significant margin. This is probably because that the language features are word embeddings trained from large-scale corpus while audio and video features are extracted manually. Thus language modality contains much richer information than other modalities.

Secondly, we compare the models with bimodal features. We can infer that when combining language modality with acoustic modality or vision modality, the performance can be improved on some metrics compared with only using language modality, but not all of them can be improved. 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 modality, suggesting that language modality is the dominated modality in this task. When cooperating three modalities, our model can achieve further improvements compared with using bimodal features.

Finally, as can be seen in Table 3, for all different combinations of modalities, the performance of the models with auxiliary task outperform that of the models without auxiliary task, which suggests that sentimental words classification auxiliary task indeed plays a remarkable role in our model. In addition, the proposed crossmodal coattention mechanism which learns the interaction between different modalities also makes significant contribution in our model. Due to the sufficient modality fusion and the cooperation of the auxiliary task, our model can achieve the final promising performance.

## 5.2 Case Study

As mentioned before, we propose a sentimental words classification task as an auxiliary task in the model to help to guide the fusion of multiple modalities and in turn help to learn more precise attention weights of sentimental words in the sentence. In order to investigate how the auxiliary task guide the learning of attention weights, we conduct a case study on two instances.

As shown in Figure 3, we present the attention weights learned by our model (SWAFN) with auxiliary task and without auxiliary task (denoted as SWAFN( $\Delta$ )). The first line of each example is the predicted labels of the word-level classification task, “N” means the word is predicted as not sentimental word, “Y” means the word is predicted as a sentimental word. For example, for sentence “And i was unbelievably shocked how much i loved it”, there are three sentimental words in this sentence which are “unbelievably”, “shocked” and “loved”, in which the word “unbelievably” and “shocked” are negative and the word “loved” is positive. The third and fourth line of each example are the learned attentions of each word by SWAFN model and SWAFN( $\Delta$ ) model. We can see that SWAFN pays most of attention

on the three sentimental words and can assign largest weight on the word which can directly reflect the sentiment of the sentence. However, SWAFN without auxiliary task (SWAFN( $\Delta$ )) pays most attention on the word “shocked”, which is a negative word, so it predicts wrong label of the sentiment. Similar observation can be seen in another instance.

The observation shown in Figure 3 indicates that the sentimental words classification auxiliary task can guide the model to pay more attention on sentimental words than other words when predicting sentiment and can recognize which sentimental words reflect the sentiment directly. With more accurate attention weights, SWAFN can summarize more effective representation, thus it can achieve better performance than SWAFN(no auxiliary task).

## 6 Conclusion

In this paper, we propose a Sentimental Words Aware Fusion Network (SWAFN) which first applies the crossmodal coattention mechanism to learn the long range of context information and then use a sentimental words classification auxiliary task to guide and learn the sentimental words aware final multimodal fusion representation. The experimental results on several datasets show the effectiveness of our model. The results and case study also demonstrate that our proposed sentimental words classification auxiliary task is an effective way to use the external knowledge to help the model to learn more powerful multimodal representation. In the future, we will consider incorporating more external language knowledge to obtain better multimodal fused representations.

## 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

- Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrusaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017*, pages 163–171.
- Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP - A collaborative voice analysis repository for speech technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014*, pages 960–964.
- Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161.
- Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pages 2247–2256.
- Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. 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*, pages 481–492.
- Navonil Majumder, Devamanyu Hazarika, Alexander F. Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl. Based Syst.*, 161:124–133.
- Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011*, pages 169–176.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543.

- Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6892–6899.
- Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 873–883.
- Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining, ICDM 2017*, pages 1033–1038.
- Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *Computer Vision - ECCV 2016 - 14th European Conference*, volume 9911 of *Lecture Notes in Computer Science*, pages 338–353.
- Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 3856–3866.
- Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. In *7th International Conference on Learning Representations, ICLR 2019*.
- Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 7216–7223.
- Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *5th International Conference on Learning Representations, ICLR 2017*.
- Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *Computing Research Repository*, arXiv:1606.06259.
- Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1103–1114.
- 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 *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5634–5641.
- Amir 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, ACL 2018*, pages 2236–2246.
- Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018c. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5642–5649.