Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis

Feiyang Chen[1], Ziqian Luo[2], Yanyan Xu[1*], and Dengfeng Ke[3**]

[1] School of Computer Science and Technology, Beijing Forestry University
[2] Language Technologies Institute, Carnegie Mellon University
[3] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
{fychen98.ai,luoziqian98}@gmail.com
xuyanyan@bjfu.edu.cn
dengfeng.ke@nlpr.ia.ac.cn

Abstract. Sentiment analysis, mostly based on text, has been rapidly developing in the last decade and has attracted widespread attention in both academia and industry. However, information in the real world usually comes from multiple modalities, such as audio and text. Therefore, in this paper, based on audio and text, we consider the task of multimodal sentiment analysis and propose a novel fusion strategy including both multi-feature fusion and multi-modality fusion to improve the accuracy of audio-text sentiment analysis. We call it the DFF-ATMF (Deep Feature Fusion - Audio and Text Modality Fusion) model, which consists of two parallel branches, the audio modality based branch and the text modality based branch. Its core mechanisms are the fusion of multiple feature vectors and multiple modality attention. Experiments on the CMU-MOSI dataset and the recently released CMU-MOSEI dataset, both collected from YouTube for sentiment analysis, show the very competitive results of our DFF-ATMF model. Furthermore, by virtue of attention weight distribution heatmaps, we also demonstrate the deep features learned by using DFF-ATMF are complementary to each other and robust. Surprisingly, DFF-ATMF also achieves new state-of-the-art results on the IEMOCAP dataset, indicating that the proposed fusion strategy also has a good generalization ability for multimodal emotion recognition.

Keywords: Multimodal Fusion· Multi-Features Fusion · Multimodal Sentiment Analysis

1 Introduction

Sentiment analysis provides beneficial information to understand an individual's attitude, behavior, and preference [33]. Understanding and analyzing context-related sentiment is an innate ability of a human being, which is also an important

---

[*] Corresponding Author

[**] Corresponding Author

distinction between a machine and a human being [11]. Therefore, sentiment analysis becomes a crucial issue in the field of artificial intelligence to be explored.

In recent years, sentiment analysis mainly focuses on textual data, and consequently, text-based sentiment analysis becomes relatively mature [33]. With the popularity of social media such as Facebook and YouTube, many users are more inclined to express their views with audio or video [21]. Audio reviews become an increasing source of consumer information and are increasingly being followed with interest by companies, researchers and consumers. They also provide more natural experiences than traditional text comments due to allowing viewers to better perceive a commentator's sentiment, belief, and intention through richer channels such as intonation [24]. The combination of multiple modalities [32,24] brings significant advantages over using only text, including language disambiguation (audio features can help eliminate ambiguous language meanings) and language sparsity (audio features can bring additional emotional information). Also, basic audio patterns can enhance links to the real world environment. Actually, people often associate information with learning and interact with the external environment through multiple modalities such as audio and text [1]. Consequently, multimodal learning becomes a new effective method for sentiment analysis [17]. Its main challenge lies in inferring joint representations that can process and connect information from multiple modalities [25].

In this paper, we propose a novel fusion strategy, including the multi-feature fusion and the multi-modality fusion, to improve the accuracy of multimodal sentiment analysis based on audio and text. We call it the DFF-ATMF model, and the learned features have strong complementarity and robustness. We conduct experiments on the CMU Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) [31] dataset and the recently released CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [32] dataset, both collected from YouTube, and make comparisons with other state-of-the-art models to show the very competitive performance of our proposed model. It is worth mentioning that DFF-ATMF also achieves the most advanced results on the IEMOCAP dataset in the generalized verification experiments, meaning that it has a good generalization ability for multimodal emotion recognition.

The major contributions of this paper are as follows:

- We propose the DFF-ATMF model for audio-text sentiment analysis, combining the multi-feature fusion with the multi-modality fusion to learn more comprehensive sentiment information.
- The features learned by the DFF-ATMF model have good complementarity and excellent robustness, and even show an amazing performance when generalized to emotion recognition tasks.
- Experimental results indicate that the proposed model outperforms the state-of-the-art models on the CMU-MOSI dataset [7] and the IEMOCAP dataset [25], and also has very competitive results on the recently released CMU-MOSEI dataset.

The rest of this paper is structured as follows. In the following section, we review related work. We exhibit the details of our proposed methodologies in

Section 3. Then, in Section 4, experimental results and further discussions are presented. Finally, we conclude this paper in Section 5.

2   Related Work

2.1   Audio Sentiment Analysis

Audio data are usually extracted from the characteristics of audio samples' channel, excitation, and prosody. Among them, prosody parameters extracted from segments, sub-segments, and hyper-segments are used for sentiment analysis in [14]. In the past several years, classical machine learning algorithms, such as Hidden Markov Model (HMM), Support Vector Machine (SVM), and decision tree-based methods, have been utilized for audio sentiment analysis [27,26,13]. Recently, researchers have proposed various neural network-based architectures to improve audio sentiment analysis. In 2014, an initial study employed deep neural networks (DNNs) to extract high-level features from raw audio data and demonstrated its effectiveness [8]. With the development of deep learning, more complex neural-based architectures have been proposed. For example, convolutional neural network (CNN)-based models have been used to train spectrograms or audio features derived from original audio signals such as Mel Frequency Cepstral Coefficients (MFCCs) and Low-Level Descriptors (LLDs) [2,20,19].

2.2   Text Sentiment Analysis

After decades of development, text sentiment analysis has become mature in recent years [9]. The most commonly used classification techniques such as SVM, maximum entropy and naive Bayes, are based on the word bag model, where the sequence of words is ignored, which may result in inefficient extraction of sentiment from the input because the sequence of words will affect the existing sentiment [4]. Later research has overcome this problem by using deep learning in sentiment analysis [33]. For instance, a kind of DNN model is proposed, using word-level, character-level and sentence-level representations for sentiment analysis [10]. In order to better capture the temporal information, [5] proposes a novel neural architecture, called Transformer-XL, which enables learning dependency beyond a fixed-length without disrupting temporal coherence. It consists of a segment-level recurrence mechanism and a novel positional encoding scheme, not only capturing longer-term dependency but also resolving the context fragmentation problem.

2.3   Multimodal Learning

Multimodal learning is an emerging field of research [1]. Learning from multiple modalities needs to capture the correlation among these modalities. Data from different modalities may have different predictive power and noise topology, with possibly losing the information of at least one of the modalities [1]. [17] presents a

novel feature fusion strategy that proceeds in a hierarchical manner for multimodal sentiment analysis. [7] proposes a recurrent neural network-based multimodal attention framework that leverages contextual information for utterance-level sentiment prediction and shows a state-of-the-art model on the CMU-MOSI and CMU-MOSEI datasets.

3   Proposed Methodology

In this section, we describe the proposed DFF-ATMF model for audio-text sentiment analysis in detail. We firstly introduce an overview of the whole neural network architecture, illustrating how to fuse audio and text modalities. After that, two separate branches of DFF-ATMF are respectively explained to show how to fuse the audio feature vector and the text feature vector. Finally, we present the multimodal-attention mechanism used in the DFF-ATMF model.
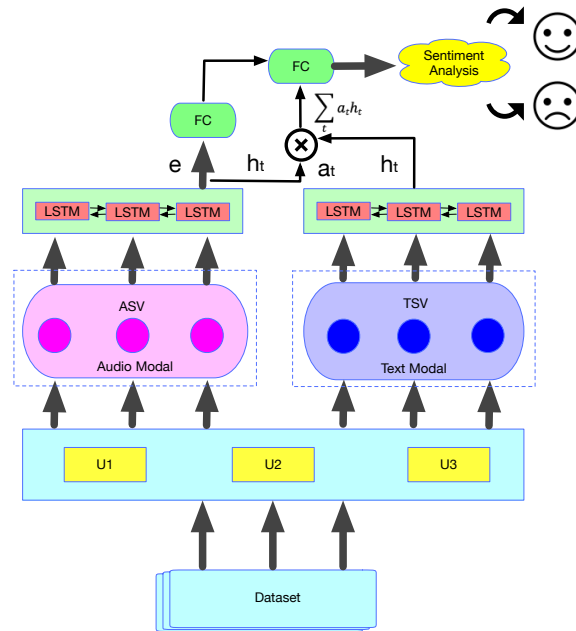

Fig. 1. The overall architecture of the proposed DFF-ATMF framework. h_t represents the hidden state of Bi-LSTM at time t. e means the final audio sentiment vector. a_t represents the attention weight and is calculated as the dot product of the final audio sentiment vector and the final text sentiment vector of h_t. "FC" means a fully-connected layer.

3.1   The DFF-ATMF Framework

The overall architecture of the proposed DFF-ATMF framework is shown in Figure 1. We fuse audio and text modalities in this framework through two parallel branches, that is, the audio modality based branch and the text modality based branch. DFF-ATMF's core mechanisms are feature vector fusion and multimodal-attention fusion. The audio modality branch uses Bi-LSTM [3] to extract audio sentiment information between adjacent utterances (U1, U2, U3), while another branch uses the same network architecture to extract text features. Furthermore, the audio feature vector of each piece of utterance is used as the input of our proposed neural network, which is based on the audio feature fusion, so we can obtain a new feature vector before the softmax layer, called the audio sentiment vector (ASV). The text sentiment vector (TSV) can be achieved similarly. Finally, after the multimodal-attention fusion, the output of the softmax layer produces final sentiment analysis results, as shown in Figure 1.

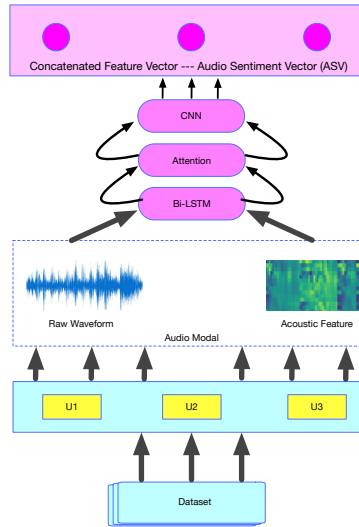### 3.2   Audio Sentiment Vector (ASV) from Audio Feature Fusion (AFF)

Fig. 2. The architecture of ASV from AFF.

Base on the work in [15], in order to explore further the fusion of feature vectors inter the audio modality, we extend the experiments of different types of audio features on the CMU-MOSI dataset, and the results are shown in Table 1.

Feature	Model	Accuracy(%)		
		2-class	5-class	7-class
1 Chromagram from spectrogram (chroma_stft)	LSTM	43.24	20.23	13.96
	BiLSTM	45.37	2.29	12.39
2 Chroma Energy Normalized (chroma_cens)	LSTM	42.98	20.87	13.31
	BiLSTM	45.85	20.53	13.76
3 Mel-frequency cepstral coefficients (MFCC)	**LSTM**	**55.12**	**23.64**	**16.99**
	BiLSTM	**55.98**	**23.75**	**17.24**
4 Root-Mean-Square Energy (RMSE)	LSTM	52.30	21.14	15.33
	BiLSTM	52.76	22.35	15.87
5 Spectral_Centroid	LSTM	48.39	22.25	14.97
	BiLSTM	48.84	22.36	15.79
6 Spectral_Contrast	LSTM	48.34	22.50	15.02
	BiLSTM	48.97	22.28	15.98
7 Tonal Centroid Features (tonnetz)	LSTM	53.78	22.67	15.83
	BiLSTM	54.24	21.87	16.01

Table 1. Comparison of different types of audio features on the CMU-MOSI dataset.

In addition, we also implement an improved serial neural network of Bi-LSTM and CNN [28], combining with the attention mechanism to learn the deep features of different sound representations. The multi-feature fusion procedure is described with the LSTM branch and the CNN branch respectively in Algorithm 1.

The features are learned from raw waveforms and acoustic features, which are complementary to each other. Therefore, audio sentiment analysis can be improved by applying our feature fusion technique, that is, ASV from AFF, whose architecture is shown in Figure 2.
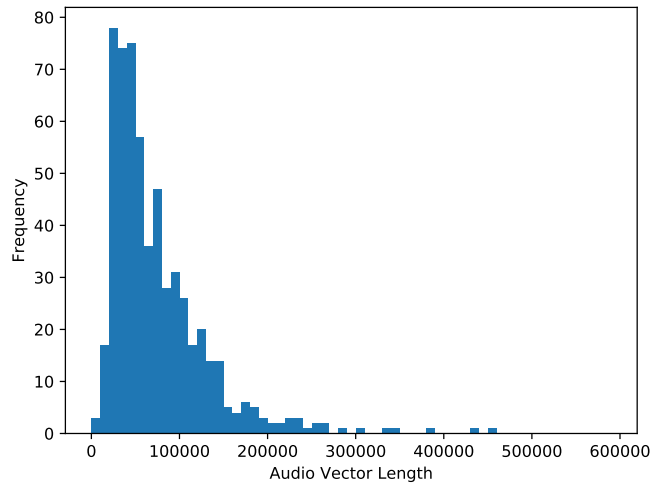

Fig. 3. The raw audio waveform sampling distribution on the CMU-MOSI dataset.

Algorithm 1 The Multi-Feature Fusion Procedure

 1: **procedure** LSTM BRANCH
 2:     **for** i:[0,n] **do**
 3:         $f_i = getAudioFeature(u_i)$ // get the audio feature from the uth utterance
 4:         $a_i = getASV(f_i)$
 5:     **end for**
 6:     **for** i:[0,M] **do** //M is the number of videos
 7:         $input_i = GetTopUtter(v_i)$
 8:         $u_{f_i} = getUtterFeature(input_i)$
 9:     **end for**
10:     $shuffle(v)$
11:     $Attention(A_i)$
12:     $Multi - Feature\ Fusion\ from\ the\ LSTM\ branch$
13: **end procedure**
14: **procedure** CNN BRANCH
15:     **for** i:[0,n] **do**
16:         $x_i \leftarrow get\ SpectrogramImage(u_i)$
17:         $c_i \leftarrow CNNModel(x_i)$
18:     **end for**
19:     $Attention(C_i)$
20:     $Multi - Feature\ Fusion\ from\ the\ CNN\ branch$
21: **end procedure**
22: **procedure** FEATURE FUSION
23:     **for** i:[0,n] **do**
24:         $L_i = Attention(a_i)$
25:         $C_i = Attention(l_i)$
26:     **end for**
27:     $Attention(L_i + C_i)$
28:     $Multi - Feature\ Fusion$
29: **end procedure**

In terms of raw audio waveforms, taking the CMU-MOSI dataset as an example, we illustrate their sampling distribution in Figure 3. The inputs to the network are raw audio waveforms sampled at 22 kHz. We also scale the waveforms to be in the range [-256, 256], so that we do not need to subtract the mean value as the data are naturally near zero already. To obtain a better sentiment analysis accuracy, batch normalization (BN) and the ReLU function are employed after each convolutional layer. Additionally, dropout regularization is also applied to the proposed serial network architecture.

In terms of acoustic features, we extract them using the Librosa [18] toolkit and obtain four effective kinds of features to represent sentiment information, which are MFCCs, spectral_centroid, chroma_stft and spectral_contrast, respectively. In particular, taking log-Mel spectrogram extraction [29] as an example, we

use 44.1 kHz without downsampling and extract the spectrograms with 64 Bin Mel-scale. The window size for short-time Fourier transform is 1,024 with a hop size of 512. The resulting Mel-spectrograms are next converted into log-scaled ones and standardized by subtracting the mean value and divided by the standard deviation.

Finally, we feed feature vectors of raw waveforms and acoustic features into our improved serial neural network of Bi-LSTM and CNN, combining with the attention mechanism to learn the deep features of different sound representations, that is, ASV.
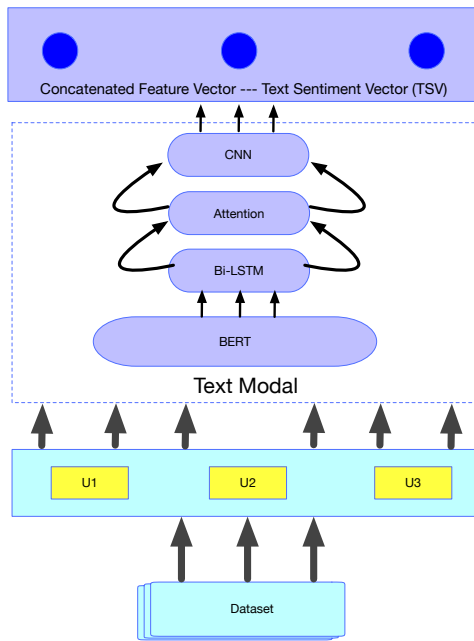

Fig. 4. The architecture of TSV from TFF.

3.3   Text Sentiment Vector (TSV) from Text Feature Fusion (TFF)

The architecture of TSV from TFF is shown in Figure 4. BERT [6] is a new language representation model, standing for Bidirectional Encoder Representations from Transformers. Thus far, to the best of our knowledge, no studies have leveraged BERT to pre-train text feature representations on the multimodal dataset such as CMU-MOSI. We then utilize BERT embeddings for CMU-MOSI. Next, the Bi-LSTM layer takes the concatenated word embeddings and POS tags

as its inputs and it outputs each hidden state. Let h_i be the output hidden state at time i. Then its attention weight a_i can be formulated by Equation 1.

$$
\begin{aligned}
m_i &= tanh(h_i) \\
\hat{a}_i &= w_i m_i + b_i \\
a_i &= \frac{exp(\hat{a}_i)}{\sum_j exp(\hat{a}_j)}
\end{aligned}
\tag{1}
$$

In Equation 1, $w_i m_i + b_i$ denotes a linear transformation of m_i. Therefore, the output representation r_i is given by:

$$
r_i = a_i h_i.
\tag{2}
$$

Based on such text representations, the sequence of features will be assigned with different attention weights. Thus, crucial information such as emotional words can be identified more easily. The convolutional layer takes the text representation r_i as its input, and the output CNN feature maps are concatenated together. Finally, text sentiment analysis can be improved by using TSV from TFF.

3.4 Audio and Text Modal Fusion with the Multimodal-Attention Mechanism

Inspired by human visual attention, the attention mechanism, proposed by [16] for neural machine translation, is introduced into the encoder-decoder framework to select reference words from the source language for the words in the target language. Based on the existing attention mechanism, inspired by the work in [30], we improve the multimodal-attention method on the basis of the multi-feature fusion strategy, focusing on the fusion of comprehensive and complementary sentiment information from audio and text. We leverage the multimodal-attention mechanism to preserve the intermediate outputs of the input sequences by retaining the Bi-LSTM encoder, and then a model is trained to selectively learn these inputs and to correlate output sequences with the model's output.

More specifically, ASV and TSV are firstly encoded with Audio-BiLSTM and Text-BiLSTM using Equation 3.

$$
\begin{aligned}
A_{t+1} &= f_\theta(A_t, x_{t+1}) \\
A_{t-1} &= f_\theta(A_t, x_{t-1}) \\
T_{t+1} &= f_\theta(T_t, x_{t+1}) \\
T_{t-1} &= f_\theta(T_t, x_{t-1})
\end{aligned}
\tag{3}
$$

In Equation 3, f_θ is the LSTM function with the weight parameter θ. A_{t+1}, A_t and A_{t-1} represent the hidden states at time $(t+1)^{th}$, t^{th} and $(t-1)^{th}$ from the audio modality, respectively. x_{t+1} and x_{t-1} represent the features at time

$(t+1)^{th}$ and $(t-1)^{th}$, respectively. The text modality is similar, represented by T.

$$a_t = \frac{exp(e^T h_t)}{\sum_t exp(e^T h_t)}$$
$$t_t = \frac{exp(e^T h'_t)}{\sum_t exp(e^T h'_t)}$$
$$Z_a = \sum_t a_t h_t \tag{4}$$
$$Z_t = \sum_t t_t h'_t$$

$$\hat{y}_{i,j} = softmax(concat(concat(Z_a, Z_t), A)^T M + b) \tag{5}$$

We then consider the final ASV e as an intermediate vector, as shown in Figure 1. During each time step t, the dot product of the intermediate vector e and the hidden state h_t is evaluated to calculate a similarity score a_t. Using this score as a weight parameter, the weighted sum $\sum_t a_t h_t$ is calculated to generate a multi-feature fusion vector Z_a. The multi-feature fusion vector of the text modality is calculated similarly, represented by Z_t. We are therefore able to obtain two kinds of multi-feature fusion vectors for the audio modality and the text modality respectively, as shown in Equation 4 and 5. These multi-feature fusion vectors are respectively concatenated with the final intermediate vectors of ASV and TSV, which will pass through the softmax function to perform sentiment analysis, as shown in Equation 6 and 7.

$$ASV = g_\theta(e)$$
$$TSV = g_\theta'(h_t) \tag{6}$$

$$\hat{y}_i = softmax(concat(ASV, TSV)^T M + b) \tag{7}$$

4   Empirical Evaluation

In this section, we firstly introduce the datasets, the evaluation metrics and the network structure parameters used in our experiments, and then exhibit the experimental results and make comparisons with other state-of-the-art models to show the advantages of DFF-ATMF. At last, more discussions are illustrated to understand the learning behavior of DFF-ATMF better.

Table 2. Datasets for training and test in our experiments.

Dataset	Training		Test	
	#utterance	#video	#utterance	#video
CMU-MOSI	1 616	65	583	28
CMU-MOSEI	18 051	1 550	4 625	679
IEMOCAP	4 290	120	1 208	31

4.1   Experiment Settings

Datasets The datasets used for training and test are depicted in Table 2. The CMU-MOSI dataset is rich in sentiment expression, consisting of 2,199 utterances, that is, 93 videos by 89 speakers. The videos involve a large array of topics such as movies, books, and other products. These videos were crawled from YouTube and segmented into utterances where each utterance is annotated with scores between -3 (strongly negative) and $+3$ (strongly positive) by five annotators. We take the average of these five annotations as the sentiment polarity and then consider only two classes, that is, "positive" and "negative". Our training and test splits of the dataset are completely disjoint with respect to speakers. In order to better compare with the previous work, similar to [25], we divide the dataset by 7:3 approximately, resulting in 1,616 and 583 utterances for training and test respectively.

The CMU-MOSEI dataset is an upgraded version of the CMU-MOSI dataset, which has 3,229 videos, that is, 22,676 utterances, from more than 1,000 online YouTube speakers. The training and test sets include 18,051 and 4,625 utterances respectively, similar to [7].

The IEMOCAP dataset was collected following theatrical theory in order to simulate natural dyadic interactions between actors. We use categorical evaluations with majority agreement and use only four emotional categories, that is, "happy", "sad", "angry", and "neutral" to compare the performance of our model with other researches using the same categories [25].

Evaluation Metrics We evaluate the performance of our proposed model by the *weighted accuracy* on 2-class or multi-class classifications.

$$weighted\ accuracy = \frac{correct\ utterances}{utterances} \tag{8}$$

Additionally, F1-Score is used to evaluate 2-class classification.

$$\mathrm{F}_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{9}$$

In Equation 9, β represents the weight between precision and recall. During our evaluation process, we set $\beta = 1$ since we consider precision and recall to have the same weight, and thus F1-score is adopted.

However, in emotion recognition, we use Macro $F1$-Score to evaluate the performance.

$$Macro\ F1 = \frac{\sum\limits_{1}^{n} F_{1n}}{n} \tag{10}$$

In Equation 10, n represents the number of classifications and F_{1n} is the $F1$ score on the nth category.

Network Structure Parameters Our proposed architecture is implemented on the open-source deep learning framework TensorFlow. More specifically, for the proposed audio and text multi-modality fusion framework, we use Bi-LSTM with 200 neurons, each followed by a dense layer consisting of 100 neurons. Utilizing the dense layer, we project the input features of audio and text to the same dimension, and next combine them with the multimodal-attention mechanism. We set the dropout hyperparameter to be 0.4 for CMU-MOSI and 0.3 for CMU-MOSEI & IEMOCAP as a measure of regularization. We also use the same dropout rates for the Bi-LSTM layers. We employ the ReLu function in the dense layers and softmax in the final classification layer. When training the network, we set the batch size to be 32, and use Adam optimizer with the cross-entropy loss function and train for 50 epochs. In data processing, we make each utterance one-to-one correspondence with the label and rename the utterance.

The network structure of the proposed audio and text multi-feature fusion framework is similar. Taking the audio multi-feature fusion framework as an example, the hidden states of Bi-LSTM are of $2 * 200$-dim. The kernel sizes of CNN are 3, 5, 7 and 9 respectively. The size of the feature map is $4 * 200$. A dropout rate is a random number between 0.3 and 0.4. The loss function used is MAE, and the batch size is set to 16. We combine the training set and the development set in our experiments. We use 90% for training and reserve 10% for cross-validation. To train the feature encoder, we follow the fine-tuning training strategy.

In order to reduce randomness and improve credibility, we report the average value over 3 runs for all experiments.

4.2   Experimental Results

Comparison with Other Models

- [22] proposes an LSTM-based model that enables utterances to capture contextual information from their surroundings in the video, thus aiding the classification.
- [23] introduces attention-based networks to improve both context learning and dynamic feature fusion.
- [32] proposes a novel multimodal fusion technique called Dynamic Fusion Graph (DFG).

Model	CMU-MOSI		CMU-MOSEI		IEMOCAP	
	Acc(%)	F1	Acc(%)	F1	Overall Acc(%)	Macro F1
[22]	79.30	80.12	-	-	75.60	76.31
[23]	80.10	80.62	-	-	-	-
[32]	74.93	75.42	76.24	77.03	-	-
[25]	76.60	76.93	-	-	78.20	78.79
[7]	80.58	80.96	79.74	80.15	-	-
[12]	-	-	**84.08**	**88.89**	-	-
DFF-ATMF	**80.98**	**81.26**	77.15	78.33	**81.37**	**82.29**

Table 3. Comparison with other state-of-the-art models.

- [25] explores three different deep learning-based architectures, each improving upon the previous one, **which is the state-of-the-art method on the IEMOCAP dataset at present.**
- [7] proposes a recurrent neural network-based multimodal-attention framework that leverages the contextual information, **which is the state-of-the-art model on the CMU-MOSI dataset at present.**
- [12] proposes a new method of learning about the hidden representations between speech and text data using CNN, **which is the state-of-the-art model on the CMU-MOSEI dataset at present.**

Table 3 shows the comparison of DFF-ATMF with other state-of-the-art models. From Table 3, we can see that DFF-ATMF outperforms the other models on the CMU-MOSI dataset and the IEMOCAP dataset. At the same time, the experimental results on the CMU-MOSEI dataset also show DFF-ATMF's competitive performance.

Generalization Ability Analysis In order to verify the feature complementarity of our proposed fusion strategy and its robustness, we conduct experiments on the IEMOCAP dataset to examine DFF-ATMF's generalization capability. Surprisingly, our proposed fusion strategy is effective on the IEMOCAP dataset and outperforms the current state-of-the-art method in [25], which can be seen from Table 3 and the overall accuracy is improved by 3.17%. More detailed experimental results on the IEMOCAP dataset are illustrated in Table 4.

4.3   Further Discussions

The above experimental results have already shown that DFF-ATMF can improve the performance of audio-text Sentiment analysis. We now analyze the attention values to understand the learning behavior of the proposed architecture better.

We take a video from the CMU-MOSI test set as an example. From the attention heatmap in Figure 5, we can see evidently that by applying different weights across contextual utterances and modalities, the model is able to predict labels of all the utterances correctly, which shows that our proposed fusion

Emotion	IEMOCAP	
	ACC(%)	Macro F1
happy	74.41	75.66
sad	73.62	74.31
angry	78.57	79.14
neutral	64.35	65.72
Overall	**81.37**	**82.29**

Table 4. Experimental results on the IEMOCAP dataset.

strategy with multi-feature and multi-modality is indeed effective, and thus has good feature complementarity and excellent robustness of generalization ability. However, at the same time, we have a doubt about the multi-feature fusion. When the raw waveform of the audio is fused with the vector of acoustic features, the dimensions are inconsistent. If the existing method is utilized to reduce the dimension, some audio information may also be lost. We intend to solve this problem from the perspective of some mathematical theory such as the angle between two vectors.
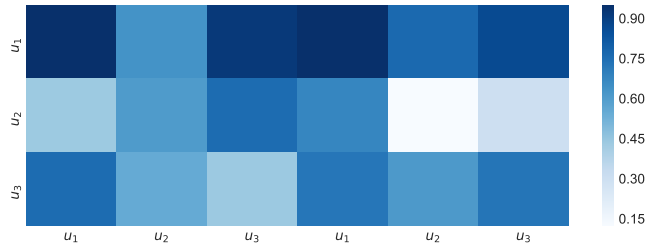

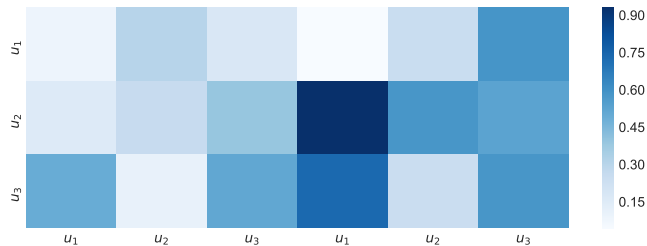**Fig. 5.** Softmax attention weights of an example from the CMU-MOSI test set.

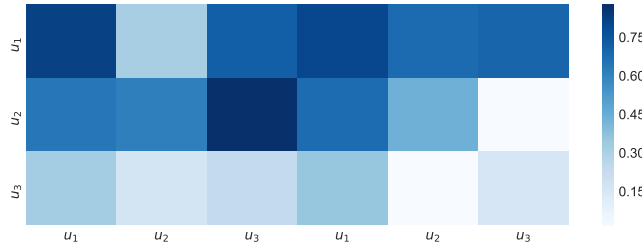**Fig. 6.** Softmax attention weights of an example from the CMU-MOSEI test set.

Fig. 7. Softmax attention weights of an example from the IEMOCAP test set.
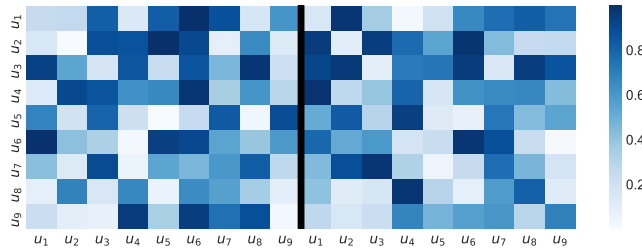

Fig. 8. Softmax attention weight comparison of the CMU-MOSI, CMU-MOSEI, and IEMOCAP test sets.

Similarly, the attention weight distribution heatmaps on the CMU-MOSEI and IEMOCAP test sets are shown in Figure 6 and 7, respectively. Furthermore, we also give the softmax attention weight comparison of the CMU-MOSI, CMU-MOSEI, and IEMOCAP test sets in Figure 8.

5   Conclusions

In this paper, we propose a novel fusion strategy, including multi-feature fusion and multi-modality fusion, and the learned features have strong complementarity and robustness, leading to the most advanced experimental results on the audio-text multimodal sentiment analysis tasks. Experiments on both the CMU-MOSI and CMU-MOSEI datasets show that our proposed model is very competitive. More surprisingly, the experiments on the IEMOCAP dataset achieve unexpected state-of-the-art results, indicating that DFF-ATMF can also be generalized for multimodal emotion recognition. In this paper, we did not consider the video modality because we try to use only the information of audio and text derived from videos. To the best of our knowledge, this is the first attempt in the multimodal domain. In the future, we will consider more fusion strategies supported by basic mathematical theories for multimodal sentiment analysis.

6   Acknowledgements

References

1. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2019)
2. Bertero, D., Fung, P.: A first look into a convolutional neural network for speech emotion detection. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5115–5119. IEEE (2017)
3. Cai, X., Hao, Z.: Multi-view and attention-based bi-lstm for weibo emotion recognition. In: 2018 International Conference on Network, Communication, Computer Engineering (NCCE), Advances in Intelligent Systems Research, volume 147. pp. 772–779. Atlantis Press (2018). https://doi.org/https://doi.org/10.2991/ncce-18.2018.127, `https://doi.org/10.2991/ncce-18.2018.127`
4. Chaturvedi, I., Cambria, E., Welsch, R.E., Herrera, F.: Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. Information Fusion **44**, 65–77 (2018)
5. Dai, Z., Yang, Z., Yang, Y., Cohen, W.W., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Ghosal, D., Akhtar, M.S., Chauhan, D., Poria, S., Ekbal, A., Bhattacharyya, P.: Contextual inter-modal attention for multi-modal sentiment analysis. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3454–3466 (2018)
8. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: The fifteenth annual conference of the international speech communication association (INTERSPEECH). pp. 223–227 (2014)
9. Hussein, D.M.E.D.M.: A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences **30**(4), 330–338 (2018)
10. Jianqiang, Z., Xiaolin, G., Xuejun, Z.: Deep convolution neural networks for twitter sentiment analysis. IEEE Access **6**, 23253–23260 (2018)
11. Kozinets, R.V., Scaraboto, D., Parmentier, M.A.: Evolving netnography: how brand auto-netnography, a netnographic sensibility, and more-than-human netnography can transform your research. JOURNAL OF MARKETING MANAGEMENT **34**(3-4), 231–242 (2018)

12. Lee, C.W., Song, K.Y., Jeong, J., Choi, W.Y.: Convolutional attention networks for multimodal emotion recognition from speech and text data. arXiv preprint arXiv:1805.06606 (2018)
13. Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. Speech Communication **53**(9-10), 1162–1171 (2011)
14. Liu, Z.T., Wu, M., Cao, W.H., Mao, J.W., Xu, J.P., Tan, G.Z.: Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing **273**, 271–280 (2018)
15. Luo, Z., Xu, H., Chen, F.: Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network. In: Proceedings of the AAAI-19 Workshop on Affective Content Analysis, Honolulu, USA, AAAI (2019)
16. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
17. Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., Poria, S.: Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-Based Systems **161**, 124–133 (2018)
18. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference. pp. 18–25 (2015)
19. Minaee, S., Abdolrashidi, A.: Deep-emotion: Facial expression recognition using attentional convolutional network. arXiv preprint arXiv:1902.01019 (2019)
20. Parthasarathy, S., Tashev, I.: Convolutional neural network techniques for speech emotion recognition. In: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). pp. 121–125. IEEE (2018)
21. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37**, 98–125 (2017)
22. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P.: Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 873–883 (2017)
23. Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A., Morency, L.P.: Multi-level multiple attentions for contextual multimodal sentiment analysis. In: 2017 IEEE International Conference on Data Mining (ICDM). pp. 1033–1038. IEEE (2017)
24. Poria, S., Hussain, A., Cambria, E.: Combining textual clues with audio-visual information for multimodal sentiment analysis. In: Multimodal Sentiment Analysis, pp. 153–178. Springer (2018)
25. Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., Hussain, A.: Multimodal sentiment analysis: Addressing key issues and setting up the baselines. IEEE Intelligent Systems **33**(6), 17–25 (2018)
26. Schuller, B., Rigoll, G., Lang, M.: Hidden markov model-based speech emotion recognition. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). vol. 2, pp. II–1. IEEE (2003)
27. Schuller, B., Rigoll, G., Lang, M.: Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 1, pp. I–577. IEEE (2004)
28. Wu, C., Wu, F., Liu, J., Yuan, Z., Wu, S., Huang, Y.: Thu_ngn at semeval-2018 task 1: Fine-grained tweet sentiment intensity analysis with attention cnn-lstm.

In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 186–192 (2018)

29. Yin, Y., Shah, R.R., Zimmermann, R.: Learning and fusing multimodal deep features for acoustic scene categorization. In: 2018 ACM Multimedia Conference on Multimedia Conference. pp. 1892–1900. ACM (2018)

30. Yoon, S., Byun, S., Jung, K.: Multimodal speech emotion recognition using audio and text. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 112–118. IEEE (2018)

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

32. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: 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). vol. 1, pp. 2236–2246 (2018)

33. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), e1253 (2018)