SPEECH-BASED AFFECTIVE COMPUTING USING ATTENTION WITH MULTIMODAL FUSION

BY YUE GU

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Electrical And Computer Engineering

Written under the direction of

Ivan Marsic

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2020

ABSTRACT OF THE DISSERTATION

Speech-based Affective Computing Using Attention With Multimodal Fusion

by Yue Gu

Dissertation Director: Ivan Marsic

Multimodal affective computing, learning to recognize and interpret human affect and subjective information from multiple data sources, is now a popular task with the recent rapid advancements in social media technology. Sentiment analysis and emotion recognition, both of which require applying subjective human concepts for detection, can be treated as two affective computing subtasks on different levels. A variety of data sources, including voice, facial expression, gesture, and linguistic content have been employed in sentiment analysis and emotion recognition. In this research, we focus on a multimodal structure to leverage the advantages of speech source on sentence-level data. Specifically, given an utterance, we consider the linguistic content and acoustic characteristics together to recognize the opinion or emotion. Our work is important and useful because speech is the most basic and commonly used form of human expression.

We first present two hybrid multimodal frameworks to predict human emotions and sentiments based on utterance-level spoken language. The hybrid deep multimodal

system extracts the high-level features from both text and audio, which considers the spatial information from text, temporal information from audio, and high-level associations from low-level handcrafted features. The system fuse all extracted features on utterance-level by using a three-layer deep neural network to learn the correlations across modalities and train the feature extraction and fusion modules together, allowing optimal global fine-tuning of the entire structure. Since not all parts of the text and vocal signals contribute equally to the predictions, a specific word may change the entire sentimental state of text; a different vocal delivery may indicate inverse emotions despite having the same linguistic content. To learn such variation, we thus introduce the hybrid attention multimodal system that consists of both feature attention and modality attention to help the model focus on learning informative representations for both modality-specific feature extraction and model fusion.

Although demonstrated for the modality attention fusion, there is still challenge to combine the textual and acoustical representations. Most previous works focused on combining multimodal information at a holistic level or fusing the extracted modality-specific features from entire utterances. However, to determine human meaning, it is critical to consider both the linguistic content of the word and how it is uttered. A loud pitch on different words may convey inverse emotions, such as the emphasis on "hell" for anger but indicating happy on "great". Synchronized attentive information on word-level across text and audio would then intuitively help recognize the sentiments and emotions. Therefore, we introduce a hierarchical multimodal architecture with attention and word-level fusion to classify utterance-level sentiment and emotion from text and audio data. Our introduced model outperforms state-of-the-art approaches on published datasets, and we demonstrate that our model's synchronized attention over

modalities offers visual interpretability.

We further propose an efficient dyadic fusion network that only relies on an attention mechanism to select representative vectors, fuse modality-specific features, and learn the sequence information. Compared to previous work, the proposed model has three distinct characteristics: 1. Instead of using a recurrent neural network to extract temporal associations as in previous research, we introduce multiple sub-view attention layers to compute the relevant dependencies among sequential utterances; this significantly improves model efficiency. 2. To improve fusion performance, we design a learnable mutual correlation factor inside each attention layer to compute associations across different modalities. 3. To overcome the label disagreement issue, we embed the labels from all annotators into a k-dimensional vector and transform the categorical problem into a regression problem; this method provides more accurate annotation information and fully uses the entire dataset. We evaluate the proposed model on two published multimodal emotion recognition datasets. Our model significantly outperforms previous state-of-the-art research by 3.8%-7.5% accuracy, using a more efficient model.

We finally introduced a novel human conversation analysis system, which uses a hierarchical encoder-decoder framework to better combine features extracted from linguistic modality, acoustic modality, and visual modality. The hierarchical structure first encodes the multimodal data into word-level features. The conversation-level encoder further selects important information from word-level features with temporal attention and represents all the conversation-level features as a vector. Considering that emotion and sentiment may change over a conversation and that multiple traits may be present simultaneously, our hierarchical decoder structure first decodes features at each time

instance. Then, the attribute decoder will further decode the feature vector at each time instance into attributes at that time.we proposed word-level fusion with modality attention. Our system achieved state-of-the-art performance on three published datasets and outperformed others at generalization testing.

Acknowledgements

I first would like to thank my advisor Prof. Ivan Marsic for his consistent support and guidance. Ivan has helped me on my research including approaching a research problem, writing a paper, and presenting my work. The meetings and conversations were vital in inspiring me to think from multiple perspectives to form a comprehensive and objective critique. I could not have completed my Ph.D. research without his guidance, advice, and support.

I would also like to sincerely thank Prof. Anand D.Sarwate, Prof. Sheng Wei, Prof. Bo Yuan, Prof. Dario Pompili, Prof. Janne Lindqvist, and Prof. Yongfeng Zhang for serving as the committee members in my Ph.D. qualification exam, thesis proposal, and thesis defense. They generously offered their time, support, and valuable suggestions for improving my research and dissertation.

I have been incredibly lucky to work with an amazing group in Ivan's lab. Thanks to: Dr. Xinyu Li, Kangning Yang, Shiyu Fu, Xinyu Lyu, Weijia Sun, Chenyang Gao, Ruiyu Zhang, Xinwei Zhao, Kaixiang Huang, Weidi Zhang, Huangcan Li, Mengzhu Li, Haotian Zhu, Chengguang Xu, and Weitian Li, who helped me during my Ph.D study. More thanks to Yanyi Zhang, Abdulbaqi Jalal, Jianyu Zhang, Sen Yang, Weizhong Kong, and Moliang Zhou, who provided feedback and suggestions through out the research projects. I would give a special thank to Shuhong Chen, who provided me a tremendous help on paper written. I would especially like to express my appreciation

to my friends. Thank you, Dr. Can Liu, Dr. Xinyu Li, Haotian Zhu, and Chengguang Xu! These extraordinary people have been my best friends and made my life colorful.

Finally, my deep and sincere gratitude to my family for their endless and unparalleled love and support. Especially thanks to my beloved wife, Hua Shang, thanks for all you support, without which I would have stopped these studys a long time ago. I would also give special thanks to my adorable cat, Xiuxiu Shang, who brings infinite happiness to my life. I really could not have accomplished all of this without them.

This work is partially supported by the National Science Foundation under Grant Number IIS-1763827. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Multimodal affective computing, learning to recognize and interpret human affect and subjective information from multiple data sources, is now a popular task with the recent rapid advancements in social media technology. Sentiment analysis and emotion recognition, both of which require applying subjective human concepts for detection, can be treated as two affective computing subtasks on different levels [Sun et al., 2017, Poria et al., 2017a]. A variety of data sources, including voice, facial expression, gesture, and linguistic content have been employed in sentiment analysis and emotion recognition [Balazs and Velásquez, 2016]. Speech based emotion recognition and sentiment analysis, which aim to automatically identify emotional or sentiment state from human verbal expression, has become an increasing expanding research topic in artificial intelligence and machine learning [El Ayadi et al., 2011, Trigeorgis et al., 2016, Badjatiya et al., 2017].

Because speech is the most basic and commonly used form of human expression [Giles and Powesland, 1975], precisely detecting human emotion or sentiment from human spoken language is useful in many real-world applications such as recommender systems and chatbots. However, it is hard for a computer to precisely interpret human affect because: 1. Giving computers the ability to detect the opinion and emotion from

speech requires a complete analysis from multiple sources such as linguistic content, vocal signals, and even need the facial expression. But how to process the heterogeneous inputs into a computer is an open-ended question. 2. It is hard to extract associated features; there is a gap between the extracted modality-specific features and the actual human affective state. The lack of high-level feature associations is a limitation of traditional approaches using low-level handcrafted features as representations. 3. Another issue is the fusion of cues from heterogeneous data. How to integrate and combine the extracted multimodal information is still challenge. In this research, we propose four different type of multimodal architectures including hybrid multimodal network, hierarchical multimodal network, mutual attentive fusion network, and human conversation analysis system to address the above issues. Specifically, given an utterance, we mainly consider the linguistic content and acoustic characteristics together to recognize the opinion or emotion. Our work focus on designing novel structures to integrate multiple source from speech data, creating effective architectures to extract the informative modality-specific features, and learning the across modality association to improve modality fusion performance.

To integrate multiple sources, we present two hybrid multimodal frameworks to predict human emotions and sentiments based on utterance-level spoken language. We first introduce the hybrid deep multimodal system to extract the high-level features from both text and audio, which considers the spatial information from text, temporal information from audio, and high-level associations from low-level handcrafted features. It uses ConvNets [Kim, 2014] to extract textual features from words and part-of-speech [Toutanova et al., 2003], a CNN-LSTM structure to capture spatial-temporal acoustic features from Mel-frequency spectral coefficients (MFSCs) energy

maps [Abdel-Hamid et al., 2014], and a three-layer deep neural network to learn high-level acoustic associations from low-level handcrafted features. We then concatenate all the extracted features by using a three-layer deep neural network to learn the mutual correlations across modalities and classify the emotions via a softmax classifier. We directly train the feature extraction module and fusion model together, so that the final loss is appropriately used to tune all parameters. The proposed structure achieves 60.4% weighted accuracy for five emotions on the IEMOCAP multimodal dataset [Busso et al., 2008]. We also demonstrate the promising performance compared with previous multimodal structures. Since not all parts of the text and vocal signals contribute equally to the predictions, a specific word may change the entire sentimental state of text; a different vocal delivery may indicate inverse emotions despite having the same linguistic content. To learn such variation, we thus introduce the hybrid attention multimodal system that consists of both feature attention and modality attention to help the model focus on learning informative representations for both modality-specific feature extraction and model fusion. To select the informative words and frames, we introduced an LSTM with an attention mechanism as the feature extractor on both the text and audio branches. A weighted pooling strategy was applied over the feature extractor to form a modality-specific feature representation. The proposed modality attention fusion overcomes the limitations from feature-level and decision-level fusion by performing feature-level fusion with modality scores over the features. We evaluated our system on three published datasets and a trauma resuscitation speech dataset. The results show that the proposed architecture achieves state-of-the-art performance. We also demonstrated the necessity of applying a multimodal structure, extracting high-level feature representations, and using modality attention fusion. The generalization

testing established that our system has the ability to handle actual speech data.

Although demonstrated for the modality attention fusion, there is still challenge to combine the textual and acoustical representations. Most previous works focused on combining multimodal information at a holistic level, such as integrating independent predictions of each modality via algebraic rules [Wöllmer et al., 2013b] or fusing the extracted modality-specific features from entire utterances [Poria et al., 2016]. They extract word-level features in a text branch, but process audio at the frame-level or utterance-level. These methods fail to properly learn the time-dependent interactions across modalities and restrict feature integration at timestamps due to the different time scales and formats of features of diverse modalities [Poria et al., 2017a]. However, to determine human meaning, it is critical to consider both the linguistic content of the word and how it is uttered. A loud pitch on different words may convey inverse emotions, such as the emphasis on "hell" for anger but indicating happy on "great". Synchronized attentive information across text and audio would then intuitively help recognize the sentiments and emotions. Therefore, we introduce a hierarchical multimodal architecture with attention and word-level fusion to classify utterance-level sentiment and emotion from text and audio data. Our model aligned the text and audio at the word-level and applied attention distributions on textual word vectors, acoustic frame vectors, and acoustic word vectors. We propose three fusion strategies with a CNN structure to combine word-level features to classify emotions. Our introduced model outperforms state-of-the-art approaches on published datasets, and we demonstrate that our model's synchronized attention over modalities offers visual interpretability.

We further propose an efficient dyadic fusion network that only relies on an attention mechanism to select representative vectors, fuse modality-specific features, and learn the sequence information. Compared to previous work, the proposed model has three distinct characteristics: 1. Instead of using a recurrent neural network to extract temporal associations as in previous research, we introduce multiple sub-view attention layers to compute the relevant dependencies among sequential utterances; this significantly improves model efficiency. 2. To improve fusion performance, we design a learnable mutual correlation factor inside each attention layer to compute associations across different modalities. 3. To overcome the label disagreement issue, we embed the labels from all annotators into a k-dimensional vector and transform the categorical problem into a regression problem; this method provides more accurate annotation information and fully uses the entire dataset. We tested our model on two published multimodal emotion recognition datasets: IEMOCAP [Busso et al., 2008] and MELD [Poria et al., 2018]. Our model shows a significant improvement in model performance and efficiency. The result indicates that our model outperforms the most recent state-of-the-art approaches by 7.5% accuracy in IEMOCAP dataset and 3.8% accuracy in MELD dataset. In addition, quantitative analysis shows the proposed modality-specific feature extraction models provide comparable results; the mutual correlation attentive factors indeed help improve fusion performance with 4.9% accuracy on IEMOCAP. We further give detailed analysis on disagreeing annotation data and provide a visualization of the inner attention.

We finally introduced a novel human conversation analysis system, which uses a hierarchical encoder-decoder framework to better combine features extracted from linguistic modality, acoustic modality, and visual modality. The hierarchical structure first

encodes the multimodal data into word-level features. The conversation-level encoder further selects important information from word-level features with temporal attention and represents all the conversation-level features as a vector. Considering that emotion and sentiment may change over a conversation and that multiple traits may be present simultaneously, our hierarchical decoder structure first decodes features at each time instance. Then, the attribute decoder will further decode the feature vector at each time instance into attributes at that time.we proposed word-level fusion with modality attention. Our system achieved state-of-the-art performance on three published datasets and outperformed others at generalization testing.

1.2 Organization

The following sections are organized as follows: We will first introduce the hybrid multimodal systems in chapter 2. In chapter 3, we are going to introduce the improved hierarchical attention multimodal network with word-level fusion strategies. In chapter 4, we propose the mutual attention fusion network. In Chapter 5 we introduce our under-going work on human conversation analysis using acoustic, textual, and visual inputs. Chapter 6 summarize our work and conclude the research.

1.3 Contribution

Our work on multimodal real-time activity recognition can be summarized :

1. Designed two hybrid multimodal networks to investigate, evaluate, and combine the low-level handcrafted features and high-level automatic generated features for the speech affective computing.

2. Introduced the hierarchical attention strategy with word-level alignment for speech emotion recognition and sentiment analysis.

3. Introduced the mutual correlation attentive factor with sub-view attention mechanism to facilitate the feature extraction and modality fusion.

4. Proposed an effective solution and a detailed experimental analysis of the label disagreement issue that keeps sequence consistency and allows full use of labeled dialog data.

5. Proposed a hierarchical encoder-decoder framework to encode acoustic, textual, and visual features from word-level to conversation-level and decode the abstract features into attribute profile at each time instance.

Chapter 2

Hybrid Multimodal Architecture

2.1 Introduction of Chapter

Human speech conveys both content and attitude. When communicating through speech, humans naturally pick up both content and emotions to understand the speaker's actual intended meaning. Emotion recognition, defined as extracting a group of affective states from humans, is necessary to automatically detect human meaning in a human-computer interaction. Speech emotion recognition, under the field of affective computing, extracts the affective states from speech and reveals the attitudes under spoken language.

Compared to the large amount of research in visual-audio multimodal emotion recognition, there is relatively little work combining text and audio modalities. To detect the emotions in utterances, humans often consider both the textual meaning and prosody. A multimodal structure is thus necessary for using both the text and audio as input data. Previous research shows promising performance improvements by combining text with acoustic information, demonstrating the potential benefits of textual-acoustic structures [Poria et al., 2015, Poria et al., 2016]. One challenge to successfully recognizing human emotions is the extraction of effective features from speech data. There are a number of widely used low-level handcrafted features used for sentiment analysis and emotion detection in natural language and speech signal processing. In particular, thousands

of low-level acoustic descriptors and derivations (LLD) with functional statistics are extracted via OpenSmile software in [Poria et al., 2016, Wöllmer et al., 2013b]; bag of words (BoW) and bag of n-grams (BoNG) were extracted from text to represent linguistic features [Schuller, 2011, Rosas et al., 2013, Jin et al., 2015]. Nevertheless, these low-level features poorly represent high-level associations and are considered insufficient to distinguish emotion [Poria et al., 2015, Poria et al., 2016, Zheng et al., 2015, Lee and Tashev, 2015]. In [Poria et al., 2015, Poria et al., 2016], a convolutional neural network (ConvNet) extracted the high-level textual features from word embedding maps to represent textual features; however, they still combined it with handcrafted low-level acoustic features in the shared representation. Although ConvNets can extract high-level acoustic features [Cai and Xia, 2015, Wang and Tashev, 2017], they do so without considering the temporal associations. Hence, a common structure that extracts high-level features from both text and audio is desirable.

Another challenge in emotion recognition is the fusion of different modalities. There are two major fusion strategies for multimodal emotion recognition: decision-level fusion and feature-level fusion. Unlike decision-level fusion that combines the unimodal results via specific rules, feature-level fusion merges the individual feature representations before the decision making, significantly improving performance [Rosas et al., 2013, Jin et al., 2015], especially in recent deep models [Poria et al., 2015, Poria et al., 2016, Gu et al., 2017b]. Nevertheless, these works directly feed the concatenated features into a classifier or use shallow-layered fusion models, which have difficulty learning the complicated mutual correlations between different modalities. A deep belief network that consists of three Restricted Boltzmann Machine layers achieves better performance than shallow fusion models by fusing the high-level audio-visual features [Zhang et al., 2017];

however, it separates the training stage of feature extraction and feature fusion. The biggest issue with this approach is that it cannot guarantee global tuning of the parameters, as the prediction loss is not actually backpropagated to tune the feature extraction module.

In this section, we first propose a deep multimodal framework to address the problems above. To predict human emotions from sentence-level spoken language, we build a hybrid deep multimodal system (**HDMS**). It uses ConvNets to extract textual features from words and part-of-speech, a CNN-LSTM structure to capture spatial-temporal acoustic features from Mel-frequency spectral coefficients (MFSCs) energy maps, and a three-layer deep neural network to learn high-level acoustic associations from low-level handcrafted features. We then concatenate all the extracted features by using a three-layer deep neural network to learn the mutual correlations across modalities and classify the emotions via a softmax classifier. We directly train the feature extraction module and fusion model together, so that the final loss is appropriately used to tune all parameters. The proposed structure achieves 60.4% weighted accuracy for five emotions on the IEMOCAP multimodal dataset. We also demonstrate the promising performance compared with previous multimodal structures.

To further improve the system performance on the feature extraction and modality fusion, we present a hybrid attention multimodal system (**HAMS**) with both feature attention and modality attention to classify utterance-level speech data. The proposed hybrid attention architecture helps the system focus on learning informative representations for both modality-specific feature extraction and model fusion. The experimental results show that our system achieves state-of-the-art or competitive results on three

published multimodal datasets. We also demonstrated the effectiveness and generalization of our system on a medical speech dataset from an actual trauma scenario. Furthermore, we provided a detailed comparison and analysis of traditional approaches and deep learning methods on both feature extraction and fusion.

2.2 Related Work

A variety of feature extraction strategies were proposed in the last decade. Early research used prosodic features to recognize human emotions [Murray and Arnott, 1993, Wu and Liang, 2010]. The vocal signal information including speaking rate, intensity, pitch, and voice quality have been introduced to form the human emotional representations [Luengo et al., 2005, Poria et al., 2017a]. Besides the prosodic features, the energy related features were demonstrated the helpfulness on the affective computing tasks. For example, the mel-frequency cepstral coefficients (MFCCs), log-frequency power coefficients (LFPCs), and linear prediction cepstral coefficients (LPCCs) were introduced as emotional features in previous work [Kishore and Satish, 2013, Kim et al., 2007, Nwe et al., 2003, Poria et al., 2017a]. Recent research proposed low-level acoustic descriptors and derivations (LLDs) with functional statistics as acoustic features [Rosas et al., 2013, Ringeval et al., 2015, Wöllmer et al., 2013a, Metallinou et al., 2012]. Different type of toolkits were applied to generate the low-level acoustic representations such as OpenSmile and COVAREP [Eyben et al., 2010b, Degottex et al., 2014b].

For textual features, the very early research rely mainly on statistical models, rule-based models, and knowledge-based models such as designing some emotional and sentimental lexicon for the specific datasets [Mishne et al., 2005, Oneto et al., 2016,

Cambria, 2016]. Most rule-based systems use bag of words (BoW) as the textual representations for the emotion or sentiment classification [Kim et al., 2000, Schuller et al., 2005, Liscombe et al., 2005]. Because the BoW representations cannot fully capture the semantic information, the the knowledge-based approaches such as bag of concepts models are introduced to provide the contextual semantic features for the sentiment and emotion classification [Wu and Tsai, 2014, Hu et al., 2013]. The basic idea of the bag of concept is to assign the words from similar classes with similar representations [Wang et al., 2014] to further improve the feature extraction and the classification. Compared with the knowledge-based approaches, statistics-based approaches also provide comparable performance in the affective computing field [Melville et al., 2009], especially in the small datasets [Socher et al., 2013].They used SVMs with bag of words (BoW) and part of speech (PoS) features in addition to low-level acoustic features [Rozgic et al., 2012, Rosas et al., 2013]. Since low-level features represent limited high-level associations [Poria et al., 2015], various deep learning approaches have been proposed in recent study, like CNNs [Poria et al., 2016] and LSTMs [Gu et al., 2017a, Zadeh et al., 2017], to learn high-level representations. To further improve system performance, an attention mechanism was introduced in machine translation and text classification [Bahdanau et al., 2014, Yang et al., 2016].

There exist two commonly used fusion strategies in previous research: decision-level fusion and feature-level fusion. Specifically, Poria et al. [Poria et al., 2015, Poria et al., 2016] used a multiple kernel learning strategy to fuse the modality data on the feature-level. A decision-level fusion was applied by Wöllmer et al. [Wöllmer et al., 2013b] that combines the results of the text and audio-visual modalities by a threshold score

vector. Deep neural network fusion was proposed in a recent study to fuse the extracted modality-specific features [Zhang et al., 2017]. More recent approaches introduced LSTM structures to fuse the features at each time step [Poria et al., 2017b, Chen et al., 2017].

2.3 Hybrid Deep Multimodal System (HDMS)

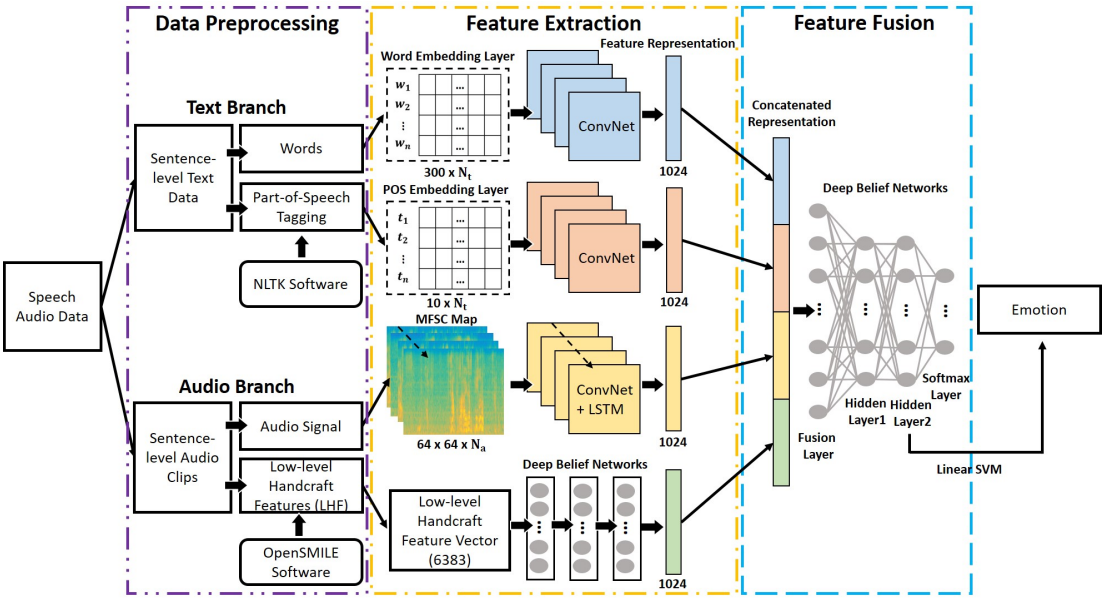


Figure 2.1: Overall structure of the proposed deep multimodal framework

2.3.1 System Overview

As shown in Fig 2.1, The proposed deep multimodal framework [1] consists of three modules: data preprocessing, feature extraction, and feature fusion. The data preprocessing module processes the input speech streams and outputs the corresponding text sentence, part-of-speech tags, audio signal, and extracted low-level handcrafted acoustic features. Then, a hybrid deep structure initializes and extracts the textual and

[1]This work has been published in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [Gu et al., 2018a].

acoustic features from the above four input branches, respectively. The fusion module concatenates the output features as a joint feature representation and learns the mutual correlations through a deep neural network. We use a softmax layer to finally predict the emotions based on the final shared representation.

2.3.2 Data Preprocessing

We first divide the input speech streams into sentence-level text and the corresponding audio clips. We used Natural Language Toolkit (NLTK) to extract the part-of-speech tags for each sentence, since phrasing also indicates the human speaking manner [Loper and Bird, 2002]. We remove all the punctuation in both the text and phrasing. Instead of just using audio signals as input data (spectral feature maps from the feature extraction module), we also extract the low-level pitch and vocal related features using OpenSmile software [Eyben et al., 2010b]. Specifically, the software extracts low-level descriptions such as fundamental frequency, pitch/energy related features, zero crossing rate (ZCR), jitter, shimmer, mel-frequency cepstral coefficients (MFCC), etc., with some functional statistics, such as flatness, skewness, quartiles, standard deviation, root quadratic mean, etc. The total number of the features is 6382. As shown in Fig 2.1, we feed all the four branches into the feature extraction module.

2.3.3 Feature Extraction

To initialize the words, we first use *word2vec* (a pre-trained word embedding model with 300 dimensions for each word based on 100 million words from Google news [Mikolov et al., 2013]) as a dictionary to embed each word into a low-dimensional word vector. We pad all sentences with zero padding to fit 40×300. As suggested

in [Kim, 2014], we apply one convolutional layer with one max-pooling layer to extract the features and use multiple convolutional filters with 2, 3, 4, and 5 as the widths. We created 256 filters for each width. The final textual feature representation is a 1024-dimensional feature vector.

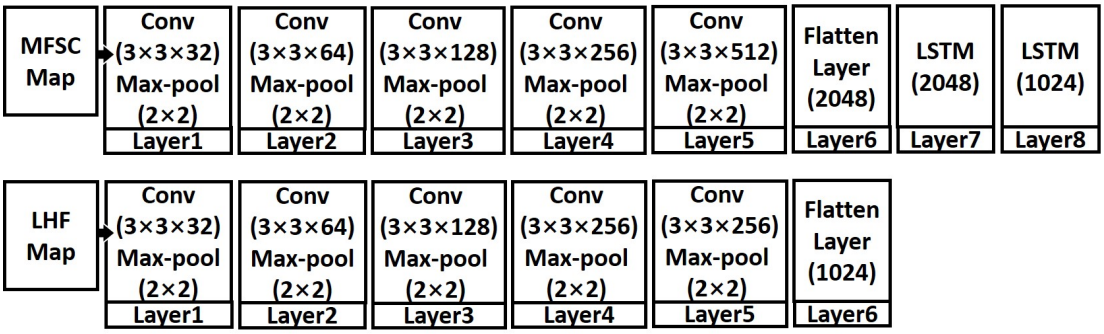


Figure 2.2: Feature extraction structure for MFSC maps

For POS embedding, we did not use a pre-trained dictionary as we did with word embedding; instead, we trained our own POS embedding dictionary based on the *word2vec* model using our own POS tagging data. We encoded the POS into a 10-dimensional vector and used the same ConvNet structure as the word branch to extract the POS features. We also created 256 filters for each width and made the output POS feature representation a 1024-dimensional feature vector.

For the audio signal input, we first extracted Mel-frequency spectral coefficients (MFSCs) from raw audio signals, which were shown to be efficient in convolutional models of speech recognition and intention classification in recent study [Gu et al., 2017b, Abdel-Hamid et al., 2014, Gu et al., 2017a]. Compared to the MFCCs, MFSCs maintain the locality of the data by preventing new basis of spectral energies resulting from discrete cosine transform in MFCC extraction [Abdel-Hamid et al., 2014]. We used 64 filter banks to extract the MFSCs and extracted both the delta and double delta coefficients. Instead of resizing the MFSC feature maps into the same size as

in [Gu et al., 2017a], we selected 64 as the context window size and 15 frames as the shift window to segment the entire MFSC map. In particular, given an audio clip, our MFSC map is a 4D array with size n×64×64×3, where n is the number of shift windows. We constructed an eight-layer ConvNet to capture the spatial associations from each MFSC segmentation, which has four convolutional layers with four max-pooling layers. As shown in Fig 2.2, we selected 3×3 as the convolutional kernel size and 2×2 as the max-pooling kernel size. We applied a fully-connected layer and a dense layer to connect feature vectors. Although previous research used a 3D-CNN structure to learn the temporal associations from the spectrograms [Zhang et al., 2017], simply concatenating output features from the ConvNet cannot reveal the actual temporal associations in sequence. LSTM is a special recurrent neural network (RNN) that allows input data with varying length, remembers values with arbitrary intervals, learns the long-term dependencies of time series, and outputs a fixed-length result. Compared with the ConvNet, LSTM is more suitable to capture the temporal associations, as it considers the sequential properties of the time series [Hochreiter and Schmidhuber, 1997]. We set up an LSTM layer after the dense layer (Layer6) to handle segmented sequential output with various lengths and learn temporal associations. We selected the hidden state from the last layer (Layer7) as the final 1024-dimensional feature vector output.

Despite the high-level acoustic features from spectral energy maps, we also extract the low-level features in prosody and vocal quality. Unlike most previous research that concatenated the low-level handcrafted features directly or reduced the dimension of the feature vectors via correlation-based feature selection (CFS) and principle component analysis (PCA) [Poria et al., 2015, Poria et al., 2016], we set up a three-layer deep

neural network of one input layer with two hidden layers to extract the high-level associations from the low-level features. Max-min normalization is applied for the low-level features before feeding them into the network. The input layer is a 6382-dimensinal feature vector and we set 2048 and 1024 as the hidden unites for each hidden layer, respectively. We select the last hidden layer as the final feature representation, which is a 1024-feature vector.

2.3.4 Feature Fusion

We concatenate all the extracted high-level features to form the joint feature representation. We use a deep neural network with one input layer, two hidden layers, and a softmax layer to capture the associations between the features from different modalities and classify the emotions [Bishop, 2006]. The hidden units are 2048 and 1024 for each hidden layer, respectively. The output of the softmax layer is the corresponding emotion vector. It worth mentioning that we also try to replace the softmax function with a linear SVM [Schuller et al., 2004] to classify the shared representation from the last hidden layer in the fusion model. Nevertheless, there is no obvious improvement in performance. To eliminate the unnecessary structures, we directly use softmax as the final classifier.

2.3.5 Network Training and Baselines

Unlike previous research that trained the feature extraction module and fusion modules separately, our architecture connects them together and uses backpropagation to adjust the entire framework, including the parameters in both fusion and feature extraction modules. Considering the multiple layers in the proposed structure, we use the rectified

linear unit (ReLU) as the activation function to facilitate convergence and set dropout functions to overcome overfitting. Another issue for training a deep model is internal covariate shift, which is defined as the change in the distribution of network activations due to the change in network parameters during training [Ioffe and Szegedy, 2015]. We applied batch normalization function between each layer to normalize and better learn the distribution [Ioffe and Szegedy, 2015], improving the training efficiency. We initialize the learning rate at 0.01 and use Adam optimizer [Kingma and Ba, 2014] to minimize the value from categorical cross-entropy loss function.

We setup the following experiment as the baselines:

- CNN_{word}: Using ConvNet as feature extractor and text as input.
- CNN_{pos}: Using ConvNet as feature extractor and part-of-speech tags as input data.
- CNN_LSTM_{mfsc}: Using CNN-LSTM as feature extractor and MFSC energy maps as input data.
- DNN_{lhaf}: Using DNN as feature extractor and low-level handcraft features as input data.
- *Both_text*: Including both CNN_{word} and CNN_{pos}.
- *Both_audio*: Including both CNN_LSTM_{mfsc} and DNN_{lhaf}.
- $LHAF_{wo}$: Low-level handcraft acoustic features without feature selection.
- $LHAF_{w}$: Low-level handcraft acoustic features with feature selection.
- CNN_{mel}: Using ConvNet as feature extractor and mel-spectrogram as input data.

Approach	Ang	Hap	Sad	Neu	Fru
CNN_{word}	42.9	54.0	50.2	39.7	49.2
CNN_{pos}	10.3	33.2	30.3	12.9	39.5
CNN_LSTM_{mfsc}	51.5	50.6	52.3	43.2	49.2
DNN_{lhaf}	54.3	44.1	40.4	39.8	41.7
$CNN_{word} + CNN_{pos}$	47.5	54.1	53.3	41.5	49.3
$CNN_{word} + CNN\_LSTM_{mfsc}$	54.6	59.2	57.2	52.1	54.3
$CNN_{word} + DNN_{lhaf}$	55.3	52.5	54.2	51.2	52.2
$CNN_{pos} + CNN\_LSTM_{mfsc}$	46.1	40.3	41.3	34.2	40.4
$CNN_{pos} + DNN_{lhaf}$	37.2	42.8	35.3	27.7	35.4
$CNN\_LSTM_{mfsc} + DNN_{lhaf}$	53.7	51.3	51.1	41.3	49.5
$Both\_text + CNN\_LSTM_{mfsc}$	55.7	61.3	57.4	52.6	57.5
$Both\_text + DNN_{lhaf}$	55.9	60.2	54.1	50.3	54.3
$CNN_{word} + Both\_audio$	56.1	63.2	60.1	55.4	60.4
$CNN_{pos} + Both\_audio$	47.2	42.3	40.1	36.2	40.5
$Ours\_separate$	55.3	61.4	57.2	52.3	58.1
$Ours\_together$	**57.2**	**65.8**	**60.2**	**56.3**	**61.6**

Table 2.1: Comparison of different feature combinations (percentage)

- CNN_{mel}: Using ConvNet as feature extractor and MFSC as input data.
- MKL: Using multiple kernel learning as fusion strategy.

2.4 Experimental Results of HDMS

We evaluate our proposed framework on the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [Busso et al., 2008]. IEMOCAP is a multimodal emotion dataset including visual, audio, and text data. In this research, we only consider the audio and text data. Three annotators assign one emotion label to each sentence from happy, sad, neutral, anger, surprised, excited, frustration, disgust, fear, and other. We only use the sentences with at least two agreed emotion labels for our experiments. We merged excited and happy as Hap, making the final dataset 1213 Hap, 1032 Sad (sad), 1084 Ang (anger), 774 Neu (neutral), and 1136 Fru (frustration). We apply 5-fold cross validation to train and test the framework.

Approach	Ang	Hap	Sad	Neu	Fru
$BoW + SVM$	40.6	45.0	42.2	31.7	44.2
CNN_{word}	42.9	54.0	50.2	39.7	49.2
$LHAF_{wo} + SVM$	41.2	36.6	38.3	39.2	41.5
$LHAF_{w} + SVM$	40.2	37.1	40.2	40.1	41.8
CNN_{mel}	39.7	41.2	43.5	39.1	41.4
$CNN_{word} + LHAF_{w} + MKL$	50.3	52.5	53.2	49.2	52.2
$CNN_{word} + CNN_{mfsc}$	50.1	52.3	56.3	51.2	50.4
$CNN_{word} + CNN_{mfsc} + SVM$	51.2	50.8	55.3	51.7	51.4
$Ours\_together$	**57.2**	**65.8**	**60.2**	**56.3**	**61.6**

Table 2.2: Comparison of previous emotion recognition structures (percentage)

We first evaluate each feature branch individually. As shown in Table 2.1, the CNN_{word} has good performance on Sad and Hap category. Compared to high-level acoustic features extracted from low-level handcrafted features (DNN_{lhaf}), the spatial-temporal high-level acoustic features extracted from the CNN-LSTM lead to better performance on Hap, Sad, Neu, and Fru. DNN_{lhaf} achieves the best result on Ang category in all unimodal structures, with 54.3% accuracy. Then, we compare the performance of different feature combinations. Combining all the features from four branches achieves the best result, with 60.4% weighted accuracy. We evaluate different training manners: training the feature extraction module and fusion module separately ($Ours\_separate$), and training all modules together ($Ours\_together$). Our result shows that training the entire structure together increases weighted accuracy by 2.7%.

We also conducted experiments using methods proposed in the previous research. From Table 2.2, our framework outperforms the text-specific model (BoW and CNN_{word}) and acoustic-specific model ($LHAF_{w}$ and CNN_{mel}) by 9.9%-29.5% accuracy. Compared with the low-level textual features (BoW), high-level textual features (CNN_{word}) improve the accuracy around 6% on average. The high-level acoustic features extracted from Mel-spectrogram via ConvNet structure (CNN_{mel}) perform slightly better than

the low-level handcrafted acoustic features without feature selection ($LHAF_{wo}$). From our result, using principal component analysis and cyclic correlation-based feature subset selection to select the low-level handcrafted acoustic features ($LHAF_w$) helps improve performance less. Both $LHAF_{wo}$ and $LHAF_w$ have lower weighted accuracies compared to DNN_{lhaf} in Table 2.1. We also evaluate structures using shallow layers in the fusion model [Poria et al., 2016, Gu et al., 2017b]; our proposed hybrid deep multimodal structure achieves the best performance, improving accuracy by up to 8%. It is worth noting that simply replacing the low-level handcrafted features with high-level features from CNN_{mfsc} in the multimodal structure does not significantly improve performance. Using CNN_LSTM_{mfsc} as the feature extractor improves 3.9% weighted accuracy, demonstrating that the lack of temporal associations indeed influences system accuracy. Our experiments also show that using a linear SVM as the classifier after the deep model does not significantly improve performance compared to a single softmax classifier.

2.5 Hybrid Attention Multimodal System (HDMS)

2.5.1 System Overview

Compared with the proposed HDMS that combines the traditional low-level features and high-level, we introduced a hybrid attention based multimodal architecture [2] for different spoken language understanding tasks. Our system used feature attention and modality attention to select the representative information at both the feature-level and modality-level. The proposed modality attention fusion overcomes the limitations from

[2]This work has been published in 2018 Proceedings of the conference. Association for Computational Linguistics. Meeting [Gu et al., 2018c].

feature-level and decision-level fusion by performing feature-level fusion with modality scores over the features. As shown in Fig 2.3, there are three major parts of the system: the data preprocessing, feature extraction, and modality fusion.

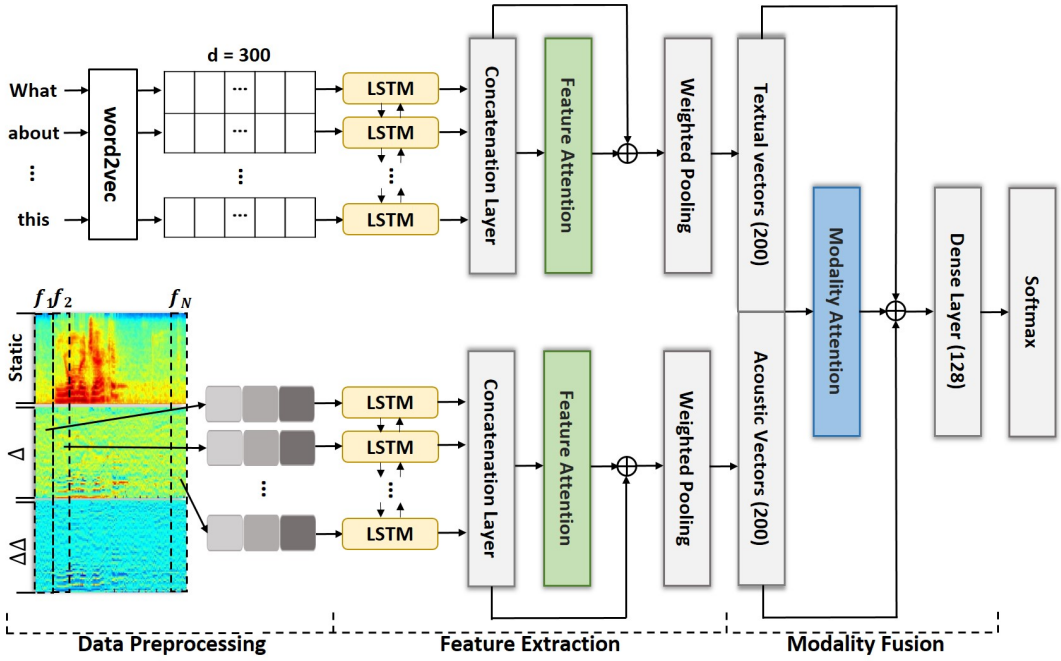


Figure 2.3: The overall system structure for hybrid attention multimodal system.

2.5.2 Data Preprocessing

The system accepts raw audio signal and text as inputs. The data preprocessing module formats the heterogeneous inputs into specific representations, which can be effectively used in the feature extraction network. We embedded the words and extracted Mel-frequency spectral coefficients (MFSCs) from the text and audio inputs for the feature extraction module. We first embedded each word into a 300-dimensional word vector by *word2vec*, which is a pre-trained word embedding dictionary trained on 100 million words from Google news [Mikolov et al., 2013]. Compared to $GloVe$ and $LexVec$,

word2vec provides us the best performance. For all embedded vectors, we allow fine-tuning of the embedding layer via backpropagation during the training stage. We removed all punctuation, as spoken language does not provide tokens. Unknown words were randomly initialized and each sentence was represented as a N $\times$300 matrix, where N is the number of the words for the given sentence. Unlike most previous research extracting LLDs or using Mel-frequency cepstral coefficients (MFCCs) as the acoustic features [Poria et al., 2016, Mirsamadi et al., 2017], we represented the raw audio signal using MFSCs because: 1. MFSCs maintain the locality of the data by preventing new bases of spectral energies resulting from discrete cosine transform in MFCCs extraction [Abdel-Hamid et al., 2014]. 2. Compared to the MFCCs that only have 39 dimensions for each audio frame, MFSCs allow more dimensions in the frequency domain that aid learning in deep models. 3. Instead of using MFCCs, voice intensity, pitch, etc. as in [Poria et al., 2017a] that need voice normalization and statistic computations, MFSC extraction does not require additional operations. As suggested in [Gu et al., 2017b], we used 64 filter banks to extract *static*, *delta*(Δ), and *doubledelta*($\Delta\Delta$) of the MFSCs as the MFSCs map. The final representation is a 3-D array with 64 $\times$F $\times$3 dimensions, where F is the number of extracted MFSCs frames.

2.5.3 Textual Feature Extraction with Attention

We applied the LSTM structure with an attention mechanism to extract temporal associations and select informative words.

The textual feature extraction module consists of two parts. Firstly, it has a regular bidirectional LSTM structure used to generate the contextual hidden states for each word vector. Secondly, it has an attention layer connected to the bidirectional LSTM to

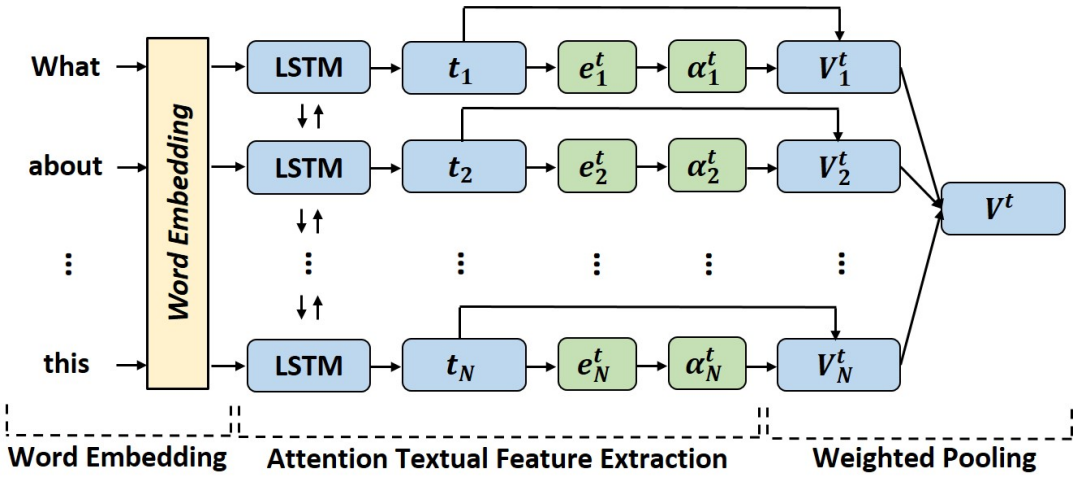


Figure 2.4: Textual feature extraction with attention.

provide a weight vector over the contextual hidden states to amplify the representative vectors. As shown in Fig 2.4, we fed the words into the bidirectional LSTM in sequence. Specifically,

$$t_i^{\rightarrow}, t_i^{\leftarrow} = bi\_LSTM(E_i), i \in [1, N] \tag{2.1}$$

where E_i is the embedded word vector of the ith word, bi_LSTM is the bidirectional LSTM, and $t_i^{\rightarrow}$ and $t_i^{\leftarrow}$ denote respectively the forward and backward contextual states of the given input word vector. Each contextual state is a word-level feature representation with forward and backward temporal associations. As not all words equally contribute to the final prediction, we added a learnable attention layer over the contextual states to denote the importance of the representations. As defined by [Bahdanau et al., 2014], we first computed the text attention energies (e_i^t) by:

$$e_i^t = tanh(W_t[t_i^{\rightarrow}, t_i^{\leftarrow}] + b_t), i \in [1, N] \tag{2.2}$$

Then, we calculated the text attention distribution (α_i^t) for word representations via a softmax function:

$$\alpha_i^t = \frac{exp({e_i^t}^\top v_t)}{\sum_{k=1}^{N} exp({e_k^t}^\top v_t)} \tag{2.3}$$

where W_t, b_t, and v_t are the learnable parameters. To form the final textual feature representation (V^t), we applied a weighted-pooling by computing a weighted sum of the text contextual states and the attention distribution:

$$V^t = \sum_{i=1}^{N} [t_i^{\rightarrow}, t_i^{\leftarrow}] \alpha_i^t \tag{2.4}$$

Unlike the systems that apply convolutional neural networks to extract the sentimental and emotional textual features using a fixed window size [Poria et al., 2015, Poria et al., 2017b], we used LSTM structures that can fully capture the sequential information with varying length and learn the temporal associations between words. We notice that Zadeh also applied LSTMs as the textual feature extractor [Zadeh et al., 2017]. However, they used a mean-pooling strategy to form the final utterance-level feature representation by passing all the contextual states into the dense layer. This assumes all the outputs can correctly contribute to the final prediction. Unfortunately, as we know, even the same word may carry diverse information that may make different contributions to the final prediction. The proposed attention layer allows the system to focus on the most informative words to further improve the representations.

2.5.4 Acoustic Feature Extraction with Attention

Similar to textual feature extraction, we also introduced a bidirectional LSTM with attention to focus on extracting informative contextual states on frame-level MFSCs.

Unlike the textual feature extraction that only has one channel (2D-array), the input MFSCs map is a 3D-array. We first concatenated the synchronized frames from static, delta, and double delta feature maps to form the input acoustic feature vector (A_j):

$$A_j = [s_j, \Delta_j, \Delta\Delta_j], j \in [1, F] \tag{2.5}$$

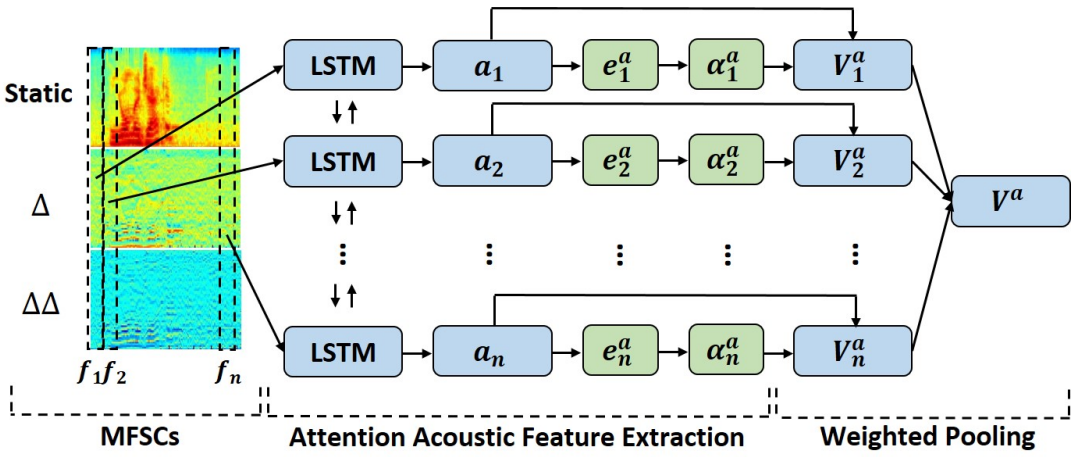


Figure 2.5: Acoustic feature extraction with attention.

Again, we used the same approach as in textual feature extraction to compute the bidirectional acoustic contextual states ($[\alpha_j^{\rightarrow}, \alpha_j^{\leftarrow}]$), acoustic attention energies (e_j^a), and acoustic attention distribution (α_j^a). The α_j^a can be understood as the importance score for the jth frame. We computed the weighted sum of the bidirectional acoustic contextual states and acoustic attention distribution as the final acoustic representation (V^a).

Unlike previous research that directly uses the acoustic LLDs as the extracted features [Degottex et al., 2014a, Poria et al., 2016], the proposed architecture learns high-level acoustic associations. We didn't use convolutional neural networks to extract the acoustic features as in [Gu et al., 2017b] because CNNs only capture spatial associations whereas acoustic data contains many temporal associations. The fixed window size of CNNs limits the temporal interaction extraction. As the number of audio frames is large (hundreds per sentence), the LSTM structure ensures the system captures long-term dependencies among the MFSCs frames. Even if a deep neural network was used for extracting the high-level associations on LLDs [Zadeh et al., 2017, Gu et al., 2018a], the generation of attention over the extracted features is still desirable, as it can help

indicate the importance at the frame-level. The weighted pooling based on the attention distribution makes sure the final acoustic feature representations contain the most informative features.

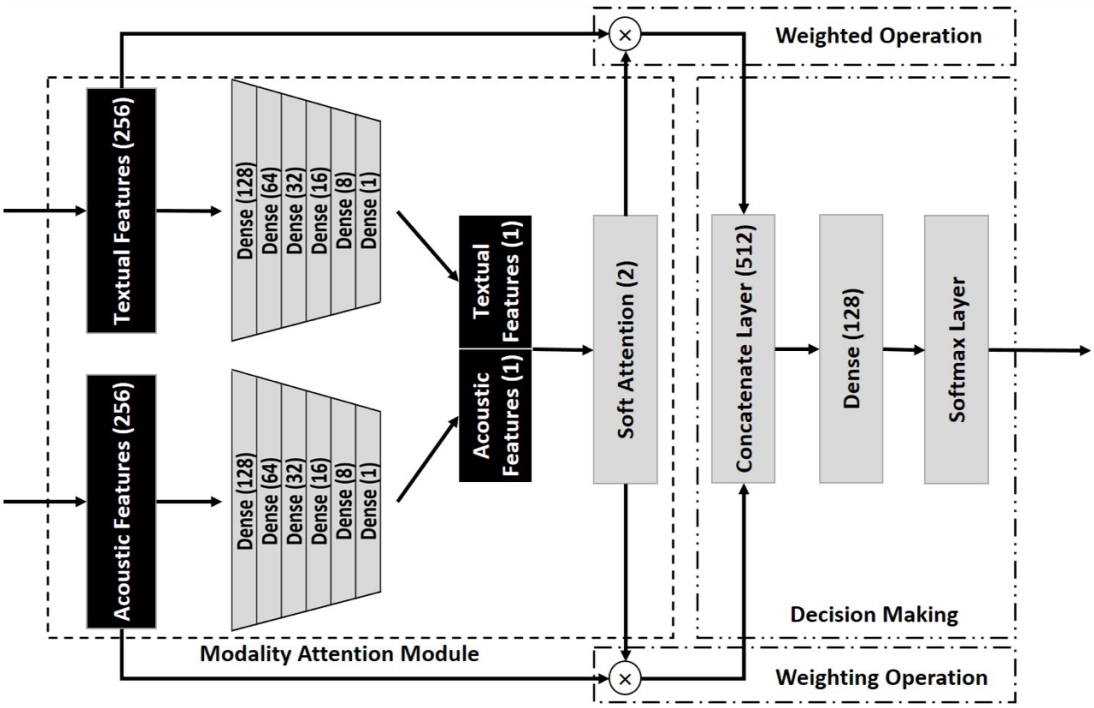


Figure 2.6: Modality fusion

2.5.5 Modality Fusion

Simply concatenating the features cannot reveal the actual importance of different modalities; the same modality may have different contributions in different spoken language understanding tasks. For example, people rely more on the vocal delivery and acoustic characteristics to express their emotions, but linguistic content and text are more important to speech content classification. Even for the same task, the modality may have distinct influences on different categories. Acoustic in-formation might provide useful information for the *anger* class, but it is hard to distinguish *neutral* and *happy* without considering text. To make the system learn this difference, we

proposed a modality attention fusion that puts an attention layer over the extracted modality-specific features, helping the system focus on the informative modality. It can be intuitively understood as giving a weighted score vector at the modality-level to indicate the importance of individual branches.

The proposed modality fusion consists of three major parts: a modality attention module, a weighted operation, and a decision making module. We first set up five dense layers after the attention layer to fuse the modality-specific features (as shown in figure 4). Then, we used softmax regression to generate the weighted score (s) for the given modality:

$$s = softmax(tanh(W_f[V^t, V^a] + b_f)) \tag{2.6}$$

where W_f and b_f are the trainable fusion attention parameters, s is a n-dimension vector, and n=2 in this study (representing the text and audio modalities respectively). We computed a soft-attention over the original modality features and concatenated them. A dense layer was used to learn the associations across weighted modality-specific features by:

$$r = tanh(W_r[(1 + s_t)V^t, (1 + s_a)V^a] + b_r) \tag{2.7}$$

where r is the final representation, and W_r and b_r are the additional parameters for the last dense layer. We used $(1 + s)$ as the attention score to keep the original modality characteristics. We made the final decision by a softmax classifier using r as input.

2.6 Experimental Results of HAMS

We evaluated the proposed system on three published multimodal datasets and an actual trauma resuscitation speech dataset. We compared our structure with the baselines from three major aspects: 1. proposed system vs previous methods; 2. low-level

Dataset	Class	Speaker Independent	Training Set	Testing Set
CMU-MOSI	2	93(74|19)	1755	444
IEMOCAP	4	151(121|30)	4295	1103
MOUD	2	79(59|20)	322	115
TRS	7	50(40|10)	7261	1843

Table 2.3: Dataset details.

handcrafted features vs high-level features; 3. shallow fusion vs deep fusion. We also conducted an experiment on a trauma resuscitation speech dataset that uses speech-to-text results as text input to test the generalizability of the system.

2.6.1 Dataset

We selected three multimodal datasets that contain spoken language information. We used audio and text data as inputs in this study. Table 2.3 shows dataset details.

CMU-MOSI: This dataset is a multimodal sentiment intensity and subjectivity dataset consisting of 93 review videos in English with 2199 utterance segments [Zadeh et al., 2016]. Each segment is labelled by five individual annotators between -3 (strong negative) to +3 (strong positive). The aim of using this dataset is to extract the sentiments from spoken language information by applying the audio segments and the corresponding transcripts. We used binary labels (positive and negative) based on the sign of the annotations' average. We used an 80-20 training-testing split that considers speaker independency. Specifically, there are 1755 utterances for training and 444 utterances for testing.

IEMOCAP: The interactive emotional dyadic motion capture database is a multimodal emotion dataset including visual, audio, and text data [Busso et al., 2008]. For this study, we only used the audio and text data and classified emotion at the utterance-level. We used the label agreed on by the majority and combined the happy and excited

classes following previous research [Poria et al., 2016]. The final dataset consists of four categories including 1591 hap (*happy+excited*), 1054 *sad*, 1076 *anger*, 1677 *neutral*. We still used an 80-20 speaker independent data split.

MOUD: The MOUD dataset is a Spanish multimodal utterance-level dataset. Following previous research [Poria et al., 2016], we only consider the positive and negative labels during training and testing. Instead of translating the sentences into English as previous research did, we initialize the word embedding layer randomly.

In addition, we tested the generalizability of the proposed system on a trauma resuscitation speech dataset (TRS).

TRS: This dataset was collected from 50 actual trauma cases with 9104 utterance-level audio segments. For each segment, it contains one utterance with at least 2 seconds. The dataset contains the following utterance-level medical category labels: *airway*, *breathing*, *circulation*, *disability*, *exposure*, $secondary-survey$, and *others*. Each utterance was assigned one category by trauma experts. The audio data was collected by two shotgun microphones placed in the resuscitation room. We used two different transcripts as the text input: human transcribed text and speech-to-text transcript. These experiments can then evaluate the influence of noise in the text branch. We reserved 40 cases as the training set and the 10 others as the testing set.

2.6.2 Baselines

We first compared our system with several state-of-the-art methods.

SVM Trees: an ensemble of SVM trees was used for classifying concatenated bag-of-words and LLDs [Rozgic et al., 2012].

BL-SVM: extracted bag-of-words and low-level descriptors as textual and acoustic

features, respectively. The model used an SVM classifier [Rosas et al., 2013].

GSV-eVector: this model used Gaussian Supervectors to select LLDs as acoustic features and extracted a set of weighted handcrafted vectors (eVector) as textual features. A linear kernel SVM was used as the final classifier [Jin et al., 2015].

C-MKL: the system used a multiple kernel learning structure as the final classifier [Poria et al., 2016]. The model extracted textual and acoustic features by using a convolution neural network and OpenSMILE software, respectively.

TFN: a tensor fusion network was used to fuse the extracted features from different modalities [Zadeh et al., 2017].

WF-LSTM: a word-level LSTM with temporal attention structure to predict sentiments on the CMU-MOSI dataset [Chen et al., 2017].

BC-LSTM: a bidirectional LSTM structure to learn contextual information among utterances [Poria et al., 2017b].

H-DMS: a hybrid deep multimodal structure to extract and fuse the textual and acoustic features on the IEMOCAP dataset [Gu et al., 2018a].

We further tested the performance of models using different feature extraction methods.

BoW: using bag-of-words as the textual features to make the final prediction [Wöllmer et al., 2013b].

WEV: directly using word embedding vectors as the textual features [Zadeh et al., 2018].

CNNs-t: Convolutional neural networks were used for extracting the textual features based on embedding word vectors [Poria et al., 2015].

LSTM-t: using an LSTM structure to learn contextual word-level textual features [Gu et al., 2017a].

OpenSimle: extracts 6373 low-level acoustic features from an entire audio clip [Poria et al., 2017b].

COVAREP: extracts low-level acoustic features including MFCCs, pitch tracking, glottal source parameters, peak slope, and maxima dispersion quotients [Chen et al., 2017].

CNNs-a: using convolutional neural networks on extracted MFSCs [Gu et al., 2017b].

LSTM-a: using an LSTM structure to learn the temporal associations based on LLDs extracted by OpenSmile [Gu et al., 2018a].

To make the comparison more reasonable, we introduced a shallow fusion and a deep fusion that combines with the previous feature extraction strategies to make the final predictions.

SVM: an SVM was trained on modality-specific features or concatenated features for classification.

DF: a deep neural network with three hidden layers was trained as the fusion module and a softmax classifier was used for decision-making.

2.6.3 Network Training

We implemented the system in Keras using the Tensorflow backend [Chollet et al., 2015, Abadi et al., 2016]. Instead of directly training the entire network, we first pre-trained the feature extraction networks by using two individual softmax classifiers. Then, we tuned the entire network by combining the feature extraction module and modality fusion module. The system was trained on a GTX 1080 GPU with 32GB RAM. We set 200 as the dimension for the bidirectional LSTM. We selected the ReLU activation function except for the attention layers. To overcome overfitting and internal covariate shift [Ioffe and Szegedy, 2015], we applied dropout and batch normalization after the

bidirectional LSTM layer and attention layers. We initialized 0.01 as the learning rate, used the Adam optimizer [Kingma and Ba, 2014], and binary/categorical cross-entropy loss. We further split 20 percent of the data from the training set as validation and used mini-batch size 8. To make a fair comparison between the proposed system and baselines, we re-trained all models on the same training-testing set split (shown in Table 2.3). We directly built the models for the baselines that provided the source code. For the rest, we re-implemented the models based on the methods described in their papers.

2.6.4 Experimental Results

We first compared the performance of the proposed system with the previous methods. The result shows that our system achieves state-of-the-art on all three published datasets. Specifically, we achieved 76.2% accuracy and 74.8 weighted F1 score on CMU-MOSI, outperforming the previous methods by a margin of 2.3% to 7.8%, which demonstrates the effectiveness of the proposed architecture. Compared to the traditional approaches using low-level handcrafted features and shallow fusion strategies (GSV-eVector and SVM Trees), the proposed method shows a significant performance improvement on IEMOCAP (9.3% and 8.7% accuracy gain, respectively). Experiments also indicate that our system performs better than the deep approaches (including C-MKL, TFN, H-DMS), showing the necessity of learning attentive information on feature extraction and fusion levels. Our approach achieves a competitive result (72.8% accuracy) on the MOUD dataset. We further re-implemented all previous methods on the TRS dataset, and our system reports the best performance in terms of both accuracy (69.4%) and weighted F1 score (66.0).

We further compare low-level vs high-level features and shallow vs deep fusion. We

	CMU-MOSI		IEMOCAP		MOUD		TRS	
Approach	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
SVM Tree	67.3	66.1	66.4	66.7	60.4	50.4	58.4	45.7
BL-SVM	68.4	67.8	65.2	65.0	60.3	52.8	59.2	50.1
GSV-eVector	65.7	65.5	64.2	64.3	61.1	52.3	58.4	48.4
C-MKL	71.3	71.0	67.0	67.2	72.0	72.2	62.1	58.1
TFN	73.6	73.5	70.4	70.2	62.1	61.2	64.4	61.5
WF-LSTM	73.9	73.3	69.5	69.4	72.7	72.8	65.6	61.5
BC-LSTM	72.4	72.6	70.8	70.8	72.4	72.4	67.9	64.4
H-DMS	70.4	70.2	70.2	69.8	68.4	67.6	66.7	64.3
Ours HAMS	**76.2**	**74.8**	**72.1**	**72.2**	**72.8**	**73.0**	**69.4**	**66.0**

Acc = accuracy (%). W-F1 = weighted accuracy.

Table 2.4: Proposed system vs previous methods.

re-trained all the individual feature extraction baselines and fusion structures on both IEMOCAP and CMU-MOSI with the same training-testing split. As shown in Table 2.5 (a), (b), and (c), we made several different combinations of the feature extraction baselines with fusion baselines. We first evaluated the performance of unimodal and multi-modal systems. From Table 2.5 (a), in all of combinations, multi-modal systems performed better than unimodal ones. In general, the performance of text is similar to that of audio on the IEMOCAP dataset, but text dominates the system performance on MOSI. This might because humans rely more on vocal delivery to express emotions, but less on sentiments. Combining textual and acoustic modalities using an ATFE+AAFE structure leads to 9.6% performance boost on IEMOCAP, which proves the necessity of using multimodal inputs in spoken language understanding. However, there is only 1.7% accuracy improvement on CMU-MOSI by using a multimodal structure. This might because humans express their attitudes without using many vocal characteristics.

Table 2.5 (b) compares the different feature extraction methods. Compared to traditional textual feature extraction (BoW), the deep models achieve better performance by extracting high-level associations on both datasets. It worth mentioning that directly

(a) Comparison of modalities

Approach	CM	IE
BoW+SVM	65.3	53.2
OS*+SVM	52.9	56.4
BoW+OS*+SVM	65.9	61.7
CNN_t+DF	69.2	57.8
CNN_a+DF	57.3	59.9
CNN_t+CNN_a+DF	71.6	64.2
ATFE+DF	74.5	61.8
AAFE+DF	60.4	62.5
ATFE+AAFE+MAF	**76.2**	**72.1**

(b) Comparison of Features

Approach	CM	IE	Approach	CM	IE
BoW+SVM	65.3	53.2	OS*+SVM	52.9	56.4
WEV+SVM	65.4	54.7	COV*+SVM	51.5	52.7
CNNt+SVM	67.3	55.2	CNN_a+SVM	54.1	55.4
LSTMt+SVM	68.2	55.7	$LSTM_a$+SVM	56.9	56.1
ATFE+SVM	72.2	61.0	AAFE+SVM	57.1	59.1
CNN_t+DF	69.2	57.8	OS*+DF	56.1	58.7
$LSTM_t$+DF	71.2	58.2	COV*+DF	55.1	56.3
$LSTM_a$+DF	58.5	60.5	CNN_a+DF	57.3	59.9
ATFE+DF	**74.5**	**61.4**	**AAFE+DF**	**60.4**	**62.5**

(c) Comparison of Fusion

Approach	CM	IE	Approach	CM	IE
BoW+OS*+SVM	65.9	61.7	CNN_t+CNN_a+SVM	65.7	63.4
BoW+OS*+DF	67.2	63.2	CNN_t+CNN_a+DF	71.6	64.2
BoW+OS*+MAF	68.7	64.7	CNN_t+CNN_a+MAF	72.9	66.1
WEA+COV*+SVM	65.8	62.7	ATFE+AAFE+SVM	71.1	65.1
WEA+COV*+DF	67.7	64.1	ATFE+AAFE+DF	74.8	70.5
WEA+COV*+MAF	68.5	64.8	**ATFE+AAFE+MAF**	**76.2**	**72.1**

(d) Generalization

Approach	TRS
AAFE+DF	56.5
ATFE(trans)+DF	66.8
ATFE(asr)+DF	47.7
ATFE(trans)+AAFE+DF	69.4
ATFE(asr)+AAFE+DF	58.9

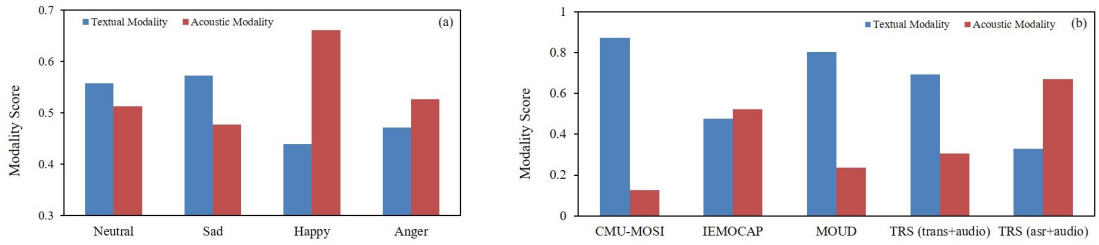
OS* = OpenSmile. COV* = COVAREP. ATFE = proposed attention based textual feature extraction. AAFE = proposed attention based acoustic feature extraction. MAF = modality attention fusion.

Table 2.5: Detailed comparison on CMU-MOSI (CM) dataset and IEMOCAP (IE) dataset (accuracy percent-age).

using the word vectors extracted by *word2vec* model as textual features (WEA+SVM) cannot outperform CNN and LSTM word vector feature extractors (CNN_t+SVM and LSTM-SVM). This observation demonstrates the necessity of extracting high-level features. On IEMOCAP, the high-level acoustic features extracted by $CNN_a$and $LSTM_a$ achieves 59.9% and 60.5% accuracy, outperforming the low-level hand-crafted acoustic features (OpenSimle+SVM and COVAREP+SVM) between 1.7% to 7.8% in accuracy. We notice that applying the LSTM architecture over the LLDs gives a 2.4% accuracy in-crease compared to directly using the LLDs on CMU-MOSI, which shows that modeling the temporal associations improve system performance. As expected, the proposed attention-based textual and acoustic feature extraction performs the best on each individual branch. Based on the above observations, we conclude that learning the high-level features from textual and acoustic data improves the system performance, and that the proposed attention-based LSTM structure indeed helps extract associated features.

Compared to the performance of shallow fusion (SVM) in Table 2.5 (c), deep fusion (DF) gives a significant performance improvement on combinations that use deep feature extractors (CNNs, LSTM, and proposed attention structure), demonstrating that extracting associations across modality-specific features indeed helps the final decision-making. The modality fusion outperforms both shallow fusion (directly using SVM classifier) and deep fusion (DF) on diverse feature ex-traction combinations. Using an MAF structure instead of SVM and DF brings 5.1% and 1.4% accuracy gain on CMU-MOSI, respectively. To further compare, we visualized the weighed scores from the modality attention on different datasets and categories (shown in Fig 2.7). We

computed the average scores of one hundred random testing samples from each category and dataset. The results indicate the proposed modality attention can learn the distinct scores on different categories and datasets.



(a) Modality attention scores of different categories on IEMOCAP. (b) Modality attention scores of different datasets.

Figure 2.7: The weighted scores of modality attention.

We further tested the generalization of the proposed system by applying it to the TRS dataset. Instead of just using the transcribed speech text, we fed the raw audio data into the IBM Watson speech to text API to automatically recognize speech (ASR). From Table 2.5 (d), using the ASR text leads to a 19.1% accuracy decrease compared to the transcribed text on unimodal systems. How-ever, the multimodal structure only has a 10.5% accuracy drop. These observations indicate that the multimodal system is tolerant to noisy data, demonstrating the generalizability of the proposed multimodal architecture with modality attention.

2.7 Summary

In this section, we first proposed a hybrid deep framework to predict the emotions from spoken language, which consists of ConvNets, CNN-LSTM, and DNN, to extract spatial and temporal associations from the raw text-audio data and low-level acoustic features. We used a four-layer deep neural network to fuse the features and classify the emotions. Our results show that the proposed framework outperforms the previous

multimodal structures on the IMOCAP dataset, achieving 60.4% weighted accuracy on five emotion categories.

we further introduced a hybrid attention based multimodal architecture for different spoken language understanding tasks. Our system used feature attention and modality attention to select the representative information at both the feature-level and modality-level. The proposed modality attention fusion overcomes the limitations from feature-level and decision-level fusion by performing feature-level fusion with modality scores over the features. We evaluated our system on three published datasets and a trauma resuscitation speech dataset. The results show that the proposed architecture achieves state-of-the-art performance. We also demonstrated the necessity of applying a multimodal structure, extracting high-level feature representations, and using modality attention fusion. The generalization testing established that our system has the ability to handle actual speech data.

Chapter 3

Hierarchical Attention Multimodal Network

3.1 Introduction of Chapter

A basic challenge in sentiment analysis and emotion recognition is filling the gap between extracted features and the actual affective states [Zhang et al., 2017]. The lack of high-level feature associations is a limitation of traditional approaches using low-level handcrafted features as representations [Seppi et al., 2008, Rozgic et al., 2012]. Recently, deep learning structures such as CNNs and LSTMs have been used to extract high-level features from text and audio [Eyben et al., 2010a, Poria et al., 2015]. However, not all parts of the text and vocal signals contribute equally to the predictions. A specific word may change the entire sentimental state of text; a different vocal delivery may indicate inverse emotions despite having the same linguistic content. Recent approaches introduce attention mechanisms to focus the models on informative words [Yang et al., 2016] and attentive audio frames [Mirsamadi et al., 2017] for each individual modality. However, to our knowledge, there is no common multimodal structure with attention for utterance-level sentiment and emotion classification. To address such issue, we design a deep hierarchical multimodal architecture [1] with an attention mechanism to classify utterance-level sentiments and emotions. It extracts high-level

[1] This work has been published in 2018 Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics Meeting [Gu et al., 2018d].

informative textual and acoustic features through individual bidirectional gated recurrent units (GRU) and uses a multi-level attention mechanism to select the informative features in both the text and audio module.

Another challenge is the fusion of cues from heterogeneous data. Most previous works focused on combining multimodal information at a holistic level, such as integrating independent predictions of each modality via algebraic rules [Wöllmer et al., 2013b] or fusing the extracted modality-specific features from entire utterances [Poria et al., 2016]. They extract word-level features in a text branch, but process audio at the frame-level or utterance-level. These methods fail to properly learn the time-dependent interactions across modalities and restrict feature integration at timestamps due to the different time scales and formats of features of diverse modalities [Poria et al., 2017a]. However, to determine human meaning, it is critical to consider both the linguistic content of the word and how it is uttered. A loud pitch on different words may convey inverse emotions, such as the emphasis on "hell" for anger but indicating happy on "great". Synchronized attentive information across text and audio would then intuitively help recognize the sentiments and emotions. Therefore, we compute a forced alignment between text and audio for each word and propose three fusion approaches (horizontal, vertical, and fine-tuning attention fusion) to integrate both the feature representations and attention at the word-level.

We evaluated our model on four published sentiment and emotion datasets. Experimental results show that the proposed architecture outperforms state-of-the-art approaches. Our methods also allow for attention visualization, which can be used for interpreting the internal attention distribution for both single- and multi-modal systems. The contributions of this paper are: (i) a hierarchical multimodal structure with

attention mechanism to learn informative features and high-level associations from both text and audio; (ii) three word-level fusion strategies to combine features and learn correlations in a common time scale across different modalities; (iii) word-level attention visualization to help human interpretation.

3.2 Related Work

Despite the large body of research on audio-visual affective analysis, there is relatively little work on combining text data. Early work combined human transcribed lexical features and low-level handcrafted acoustic features using feature-level fusion [Forbes-Riley and Litman, 2004, Litman and Forbes-Riley, 2004]. Others used SVMs fed bag of words (BoW) and part of speech (POS) features in addition to low-level acoustic features [Seppi et al., 2008, Rozgic et al., 2012, Savran et al., 2012, Rosas et al., 2013, Jin et al., 2015]. All of the above extracted low-level features from each modality separately. More recently, deep learning was used to extract higher-level multimodal features. Bidirectional LSTMs were used to learn long-range dependencies from low-level acoustic descriptors and derivations (LLDs) and visual features [Eyben et al., 2010a, Wöllmer et al., 2013b]. CNNs can extract both textual [Poria et al., 2015] and visual features [Poria et al., 2016] for multiple kernel learning of feature-fusion. Later, hierarchical LSTMs were used [Poria et al., 2017b]. A deep neural network was used for feature-level fusion in [Gu et al., 2018a] and [Zadeh et al., 2017] introduced a tensor fusion network to further improve the performance. A very recent work using word-level fusion was provided by [Chen et al., 2017]. The key differences between this work and the proposed architecture are: (i) we design a fine-tunable hierarchical attention structure to extract word-level features for each individual modality, rather than

simply using the initialized textual embedding and extracted LLDs from COVAREP [Degottex et al., 2014b]; (ii) we propose diverse representation fusion strategies to combine both the word-level representations and attention weights, instead of using only word-level fusion; (iii) our model allows visualizing the attention distribution at both the individual modality and at fusion to help model interpretability.

Our architecture is inspired by the document classification hierarchical attention structure that works at both the sentence and word level [Yang et al., 2016]. For audio, an attention-based BLSTM and CNN were applied to discovering emotion from frames [Huang and Narayanan, 2016, Neumann and Vu, 2017]. Frame-level weighted-pooling with local attention was shown to outperform frame-wise, final-frame, and frame-level mean-pooling for speech emotion recognition [Mirsamadi et al., 2017].

3.3 Methodology

We introduce a multimodal hierarchical attention structure with word-level alignment for sentiment analysis and emotion recognition (Fig. 3.1). The model consists of three major parts: text attention module, audio attention module, and word-level fusion module. We first make a forced alignment between the text and audio during preprocessing. Then, the text attention module and audio attention module extract the features from the corresponding inputs (shown in Algorithm 1). The word-level fusion module fuses the extracted feature vectors and makes the final prediction via a shared representation (shown in Algorithm 2).

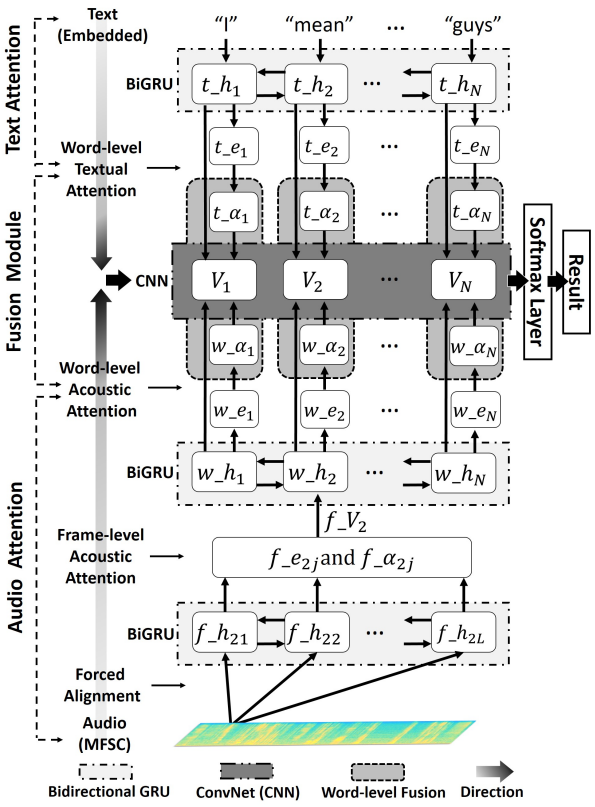


Figure 3.1: The overall system structure for multimodal hierarchical attention structure with word-level alignment.

3.3.1 Forced Alignment and Preprocessing

The forced alignment between the audio and text on the word-level prepares the different data for feature extraction. We align the data at the word-level because words are the basic unit in English for human speech comprehension. We used *aeneas*[2] to determine the time interval for each word in the audio file based on the Sakoe-Chiba Band Dynamic Time Warping (DTW) algorithm [Sakoe and Chiba, 1978].

For the text input, we first embedded the words into 300-dimensional vectors by *word2vec* [Mikolov et al., 2013], which gives us the best result compared to GloVe and LexVec. Unknown words were randomly initialized. Given a sentence S with N words,

[2]https://www.readbeyond.it/aeneas/

let w_i represent the ith word. We embed the words through the *word2vec* embedding matrix W_e by:

$$T_i = W_e w_i, i \in [1, N] \tag{3.1}$$

where T_i is the embedded word vector.

For the audio input, we extracted Mel-frequency spectral coefficients (MFSCs) from raw audio signals as acoustic inputs for two reasons. Firstly, MFSCs maintain the locality of the data by preventing new bases of spectral energies resulting from discrete cosine transform in MFCCs extraction [Abdel-Hamid et al., 2014]. Secondly, it has more dimensions in the frequency domain that aid learning in deep models [Gu et al., 2017b]. We used 64 filter banks to extract the MFSCs for each audio frame to form the MFSCs map. To facilitate training, we only used static coefficients. Each word's MFSCs can be represented as a matrix with $64 \times n$ dimensions, where n is the interval for the given word in frames. We zero-pad all intervals to the same length L, the maximum frame numbers of the word in the dataset. We did extract LLD features using OpenSmile [Eyben et al., 2010b] software and combined them with the MFSCs during our training stage. However, we did not find an obvious performance improvement, especially for the sentiment analysis. Considering the training cost of the proposed hierarchical acoustic architecture, we decided the extra features were not worth the tradeoff. The output is a 3D MFSCs map with dimensions $[N, 64, L]$.

3.3.2 Text Attention Module

To extract features from embedded text input at the word level, we first used bidirectional GRUs, which are able to capture the contextual information between words. It

Algorithm 1 FEATURE EXTRACTION

procedure FORCED ALIGNMENT
    Determine time interval of each word
    find $w_i \leftarrow \rightarrow [A_{ij}]$, $j \in [1, L]$, $i \in [1, N]$
end procedure
procedure TEXT BRANCH
    Text Attention Module
    for $i \in [1, N]$ **do**
        $T_i \leftarrow getEmbedded(w_i)$
        $t\_h_i \leftarrow bi\_GRU(T_i)$
        $t\_e_i \leftarrow getEnergies(t\_h_i)$
        $t\_\alpha_i \leftarrow getDistribution(t\_e_i)$
    end for
    return t_h_i, t_α_i
end procedure
procedure AUDIO BRANCH
    for $i \in [1, N]$ **do**
        Frame-Level Attention Module
        for $j \in [1, L]$ **do**
            $f\_h_{ij} \leftarrow bi\_GRU(A_{ij})$
            $f\_e_{ij} \leftarrow getEnergies(f\_h_{ij})$
            $f\_\alpha_{ij} \leftarrow getDistribution(f\_e_{ij})$
        end for
        $f\_V_i \leftarrow weightedSum(f\_\alpha_{ij}, f\_h_{ij})$
        Word-Level Attention Module
        $w\_h_i \leftarrow bi\_GRU(f\_V_i)$
        $w\_e_i \leftarrow getEnergies(w\_h_i)$
        $w\_\alpha_i \leftarrow getDistribution(w\_e_i)$
    end for
    return w_h_i, w_α_i
end procedure

can be represented as:

$$t\_h_i^{\rightarrow}, t\_h_i^{\leftarrow} = bi\_GRU(T_i), i \in [1, N] \tag{3.2}$$

where bi_GRU is the bidirectional GRU, $t\_h_i^{\rightarrow}$ and $t\_h_i^{\leftarrow}$ denote respectively the forward and backward contextual state of the input text. We combined $t\_h_i^{\rightarrow}$ and $t\_h_i^{\leftarrow}$ as t_h_i to represent the feature vector for the ith word. We choose GRUs instead of LSTMs because our experiments show that LSTMs lead to similar performance (0.07% higher accuracy) with around 25% more trainable parameters.

To create an informative word representation, we adopted a word-level attention strategy that generates a one-dimensional vector denoting the importance for each word in a sequence [Yang et al., 2016]. As defined by [Bahdanau et al., 2014], we compute the textual attentive energies t_e_i and textual attention distribution t_α_i by:

$$t\_e_i = tanh(W_t t\_h_i + b_t), i \in [1, N] \tag{3.3}$$

$$t\_\alpha_i = \frac{exp(t\_e_i{}^\top v_t)}{\sum_{k=1}^{N} exp(t\_e_k{}^\top v_t)} \tag{3.4}$$

where W_t and b_t are the trainable parameters and v_t is a randomly-initialized word-level weight vector in the text branch. To learn the word-level interactions across modalities, we directly use the textual attention distribution t_α_i and textual bidirectional contextual state t_h_i as the output to aid word-level fusion, which allows further computations between text and audio branch on both the contextual states and attention distributions.

3.3.3 Audio Attention Module

We designed a hierarchical attention model with frame-level acoustic attention and word-level attention for acoustic feature extraction.

Frame-level Attention captures the important MFSC frames from the given word to generate the word-level acoustic vector. Similar to the text attention module, we used a bidirectional GRU:

$$f\_h_{ij}^{\rightarrow}, f\_h_{ij}^{\leftarrow} = bi\_GRU(A_{ij}), j \in [1, L] \tag{3.5}$$

where $f\_h_{ij}^{\rightarrow}$ and $f\_h_{ij}^{\leftarrow}$ denote the forward and backward contextual states of acoustic frames. A_{ij} denotes the MFSCs of the jth frame from the ith word, $i \in [1, N]$. f_h_{ij} represents the hidden state of the jth frame of the ith word, which consists of $f\_h_{ij}^{\rightarrow}$

and $f\_h_{ij}^{\leftarrow}$. We apply the same attention mechanism used for textual attention module to extract the informative frames using equation 3.3 and 3.4. As shown in Figure **??**, the input of equation 3.3 is f_h_{ij} and the output is the frame-level acoustic attentive energies f_e_{ij}. We calculate the frame-level attention distribution f_α_{ij} by using f_e_{ij} as the input for equation 3.4. We form the word-level acoustic vector f_V_i by taking a weighted sum of bidirectional contextual state f_h_{ij} of the frame and the corresponding frame-level attention distribution f_α_{ij} Specifically,

$$f\_V_i = \sum\nolimits_j f\_\alpha_{ij} f\_h_{ij} \tag{3.6}$$

Word-level Attention aims to capture the word-level acoustic attention distribution w_α_i based on formed word vector f_V_i. We first used equation 3.2 to generate the word-level acoustic contextual states w_h_i, where the input is f_V_i and $w\_h_i = (w\_h_i^{\rightarrow}, w\_h_i^{\leftarrow})$. Then, we compute the word-level acoustic attentive energies w_e_i via equation 3.3 as the input for equation 3.4. The final output is an acoustic attention distribution w_α_i from equation 3.4 and acoustic bidirectional contextual state w_h_i.

3.3.4 Word-level Fusion Module

Fusion is critical to leveraging multimodal features for decision-making. Simple feature concatenation without considering the time scales ignores the associations across modalities. We introduce word-level fusion capable of associating the text and audio at each word. We propose three fusion strategies (Figs. 3.2 and Algorithm 2): horizontal fusion, vertical fusion, and fine-tuning attention fusion. These methods allow easy synchronization between modalities, taking advantage of the attentive associations across text and audio, creating a shared high-level representation.

Algorithm 2 FUSION

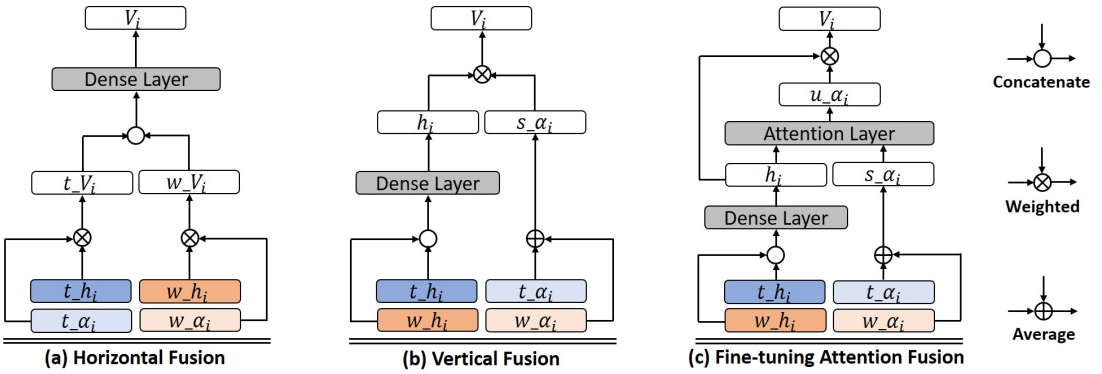
procedure FUSION BRANCH
    Horizontal Fusion (HF)
    for $i \in [1, N]$ **do**
        $t\_V_i \leftarrow weighted(t\_\alpha_i, t\_h_i)$
        $w\_V_i \leftarrow weighted(w\_\alpha_i, w\_h_i)$
        $V_i \leftarrow dense([t\_V_i, w\_V_i])$
    end for
    Vertical Fusion (VF)
    for $i \in [1, N]$ **do**
        $h_i \leftarrow dense([t\_h_i, w\_h_i])$
        $s\_\alpha_i \leftarrow average([t\_\alpha_i, w\_\alpha_i])$
        $V_i \leftarrow weighted(h_i, s\_\alpha_i)$
    end for
    Fine-tuning Attention Fusion (FAF)
    for $i \in [1, N]$ **do**
        $u\_e_i \leftarrow getEnergies(h_i)$
        $u\_\alpha_i \leftarrow getDistribution(u\_e_i, s\_\alpha_i)$
        $V_i \leftarrow weighted(h_i, u\_\alpha_i)$
    end for
    Decision Making
    $E \leftarrow convNet(V_1, V_2, ..., V_N)$
    return E
end procedure

Horizontal Fusion (HF) provides the shared representation that contains both the textual and acoustic information for a given word (Figure 3.2 (a)). The HF has two steps: (i) combining the bidirectional contextual states (t_h_i and w_h_i in Figure 3.2) and attention distributions for each branch (t_α_i and w_α_i in Figure 3.2) independently to form the word-level textual and acoustic representations. As shown in Figure 3.2, given the input (t_α_i, t_h_i) and (w_α_i, w_h_i), we first weighed each input branch by:

$$t\_V_i = t\_\alpha_i t\_h_i \tag{3.7}$$

$$w\_V_i = w\_\alpha_i w\_h_i \tag{3.8}$$

where t_V_i and w_V_i are word-level representations for text and audio branches, respectively; (ii) concatenating them into a single space and further applying a dense layer to create the shared context vector V_i, and $V_i = (t\_V_i, w\_V_i)$. The HF combines



t_h_i: word-level textual bidirectional state. t_α_i: word-level textual attention distribution. w_h_i: word-level acoustic bidirectional state. w_α_i: word-level acoustic attention distribution. s_α_i: shared attention distribution. u_α_i: fine-tuning attention distribution. V_i: shared word-level representation.

Figure 3.2: Fusion strategies.

the unimodal contextual states and attention weights; there is no attention interaction between the text modality and audio modality. The shared vectors retain the most significant characteristics from respective branches and encourages the decision making to focus on local informative features.

Vertical Fusion (VF) combines textual attentions and acoustic attentions at the word-level, using a shared attention distribution over both modalities instead of focusing on local informative representations (Figure 3.2 (b)). The VF is computed in three steps: (i) using a dense layer after the concatenation of the word-level textual (t_h_i) and acoustic (w_h_i) bidirectional contextual states to form the shared contextual state h_i; (ii) averaging the textual (t_α_i) and acoustic (w_α_i) attentions for each word as the shared attention distribution s_α_i; (iii) computing the weight of h_i and s_α_i as final shared context vectors V_i, where $V_i = h_i s\_\alpha_i$. Because the shared attention distribution (s_α_i) is based on averages of unimodal attentions, it is a joint attention of both textual and acoustic attentive information.

Fine-tuning Attention Fusion (FAF) preserves the original unimodal attentions

and provides a fine-tuning attention for the final prediction (Figure3.2 (c)). The averaging of attention weights in vertical fusion potentially limits the representational power. Addressing such issue, we propose a trainable attention layer to tune the shared attention in three steps: (i) computing the shared attention distribution s_α_i and shared bidirectional contextual states h_i separately using the same approach as in vertical fusion; (ii) applying attention fine-tuning:

$$u\_e_i = tanh(W_u h_i + b_u) \tag{3.9}$$

$$u\_\alpha_i = \frac{exp(u\_e_i{}^\top v_u)}{\sum_{k=1}^{N} exp(u\_e_k{}^\top v_u)} + s\_\alpha_i \tag{3.10}$$

where W_u, b_u, and v_u are additional trainable parameters. The u_α_i can be understood as the sum of the fine-tuning score and the original shared attention distribution s_α_i; (iii) calculating the weight of u_α_i and h_i to form the final shared context vector V_i.

3.3.5 Decision Making

The output of the fusion layer V_i is the ith shared word-level vectors. To further make use of the combined features for classification, we applied a CNN structure with one convolutional layer and one max-pooling layer to extract the final representation from shared word-level vectors [Poria et al., 2016, Wang et al., 2016]. We set up various widths for the convolutional filters [Kim, 2014] and generated a feature map c_k by:

$$f_i = tanh(W_c V_{i:i+k-1} + b_c) \tag{3.11}$$

$$c_k = max\{f_1, f_2, ..., f_N\} \tag{3.12}$$

where k is the width of the convolutional filters, f_i represents the features from window i to $i + k - 1$. W_c and b_c are the trainable weights and biases. We get the final

representation c by concatenating all the feature maps. A softmax function is used for the final classification.

3.4 Experiments

3.4.1 Datasets

We evaluated our model on four published datasets: two multimodal sentiment datasets (MOSI and YouTube) and two multimodal emotion recognition datasets (IEMOCAP and EmotiW).

MOSI dataset is a multimodal sentiment intensity and subjectivity dataset consisting of 93 reviews with 2199 utterance segments [Zadeh et al., 2016]. Each segment was labeled by five individual annotators between -3 (strong negative) to +3 (strong positive). We used binary labels based on the sign of the annotations' average.

YouTube dataset is an English multimodal dataset that contains 262 positive, 212 negative, and 133 neutral utterance-level clips provided by [Morency et al., 2011]. We only consider the positive and negative labels during our experiments.

IEMOCAP is a multimodal emotion dataset including visual, audio, and text data [Busso et al., 2008]. For each sentence, we used the label agreed on by the majority (at least two of the three annotators). In this study, we evaluate both the 4-catgeory (*happy+excited*, *sad*, *anger*, and *neutral*) and 5-catgeory(*happy+excited*, *sad*, *anger*, *neutral*, and *frustration*) emotion classification problems. The final dataset consists of 586 *happy*, 1005 *excited*, 1054 *sad*, 1076 *anger*, 1677 *neutral*, and 1806 *frustration*.

EmotiW[3] is an audio-visual multimodal utterance-level emotion recognition dataset

[3]https://cs.anu.edu.au/few/ChallengeDetails.html

consist of video clips. To keep the consistency with the IEMOCAP dataset, we used four emotion categories as the final dataset including 150 *happy*, 117 *sad*, 133 *anger*, and 144 *neutral*. We used IBM Watson[4] speech to text software to transcribe the audio data into text.

3.4.2 Baselines

We compared the proposed architecture to published models. Because our model focuses on extracting sentiment and emotions from human speech, we only considered the audio and text branch applied in the previous studies.

Sentiment Analysis Baselines

BL-SVM extracts a bag-of-words as textual features and low-level descriptors as acoustic features. An SVM structure is used to classify the sentiments [Rosas et al., 2013].

LSTM-SVM uses LLDs as acoustic features and bag-of-n-grams (BoNGs) as textual features. The final estimate is based on decision-level fusion of text and audio predictions [Wöllmer et al., 2013b].

C-MKL$_1$ uses a CNN structure to capture the textual features and fuses them via multiple kernel learning for sentiment analysis [Poria et al., 2015].

TFN uses a tensor fusion network to extract interactions between different modality-specific features [Zadeh et al., 2017].

LSTM(A) introduces a word-level LSTM with temporal attention structure to predict sentiments on MOSI dataset [Chen et al., 2017].

[4]https://www.ibm.com/watson/developercloud/speech-to-text/api/v1/

Emotion Recognition Baselines

SVM Trees extracts LLDs and handcrafted bag-of-words as features. The model automatically generates an ensemble of SVM trees for emotion classification [Rozgic et al., 2012].

GSV-eVector generates new acoustic representations from selected LLDs using Gaussian Supervectors and extracts a set of weighed handcrafted textual features as an eVector. A linear kernel SVM is used as the final classifier [Jin et al., 2015].

C-MKL$_2$ extracts textual features using a CNN and uses openSMILE to extract 6373 acoustic features. Multiple kernel learning is used as the final classifier [Poria et al., 2016].

H-DMS uses a hybrid deep multimodal structure to extract both the text and audio emotional features. A deep neural network is used for feature-level fusion [Gu et al., 2018a].

Fusion Baselines

Utterance-level Fusion (UL-Fusion) focuses on fusing text and audio features from an entire utterance [Gu et al., 2017b]. We simply concatenate the textual and acoustic representations into a joint feature representation. A softmax function is used for sentiment and emotion classification.

Decision-level Fusion (DL-Fusion) Inspired by [Wöllmer et al., 2013b], we extract textual and acoustic sentence representations individually and infer the results via two softmax classifiers, respectively. As suggested by Wöllmer, we calculate a weighted sum of the text (1.2) result and audio (0.8) result as the final prediction.

3.4.3 Model Training

We implemented the model in Keras with Tensorflow as the backend. We set 100 as the dimension for each GRU, meaning the bidirectional GRU dimension is 200. For

the decision making, we selected 2, 3, 4, and 5 as the filter width and apply 300 filters for each width. We used the rectified linear unit (ReLU) activation function and set 0.5 as the dropout rate. We also applied batch normalization functions between each layer to overcome internal covariate shift [Ioffe and Szegedy, 2015]. We first trained the text attention module and audio attention module individually. Then, we tuned the fusion network based on the word-level representation outputs from each fine-tuning module. For all training procedures, we set the learning rate to 0.001 and used Adam optimization and categorical cross-entropy loss. For all datasets, we considered the speakers independent and used an 80-20 training-testing split. We further separated 20% from the training dataset for validation. We trained the model with 5-fold cross validation and used 8 as the mini batch size. We set the same amount of samples from each class to balance the training dataset during each iteration.

3.5 Result Analysis

3.5.1 Comparison with Baselines

The experimental results of different datasets show that our proposed architecture achieves state-of-the-art performance in both sentiment analysis and emotion recognition (Table 3.1). We re-implemented some published methods [Rosas et al., 2013, Wöllmer et al., 2013b] on MOSI to get baselines.

For sentiment analysis, the proposed architecture with FAF strategy achieves 76.4% weighted accuracy, which outperforms all the five baselines (Table 3.1). The result demonstrates that the proposed hierarchical attention architecture and word-level fusion strategies indeed help improve the performance. There are several findings worth

Sentiment Analysis (MOSI)					Emotion Recognition (IEMOCAP)				
Approach	Category	WA(%)	UA(%)	F1	Approach	Category	WA(%)	UA(%)	F1
BL-SVM*	2-class	70.4	70.6	0.668	SVM Trees	4-class	67.4	67.4	-
LSTM-SVM*	2-class	72.1	72.1	0.674	GSV-e Vector	4-class	63.2	62.3	-
C-MKL$_1$	2-class	73.6	-	0.752	C-MKL$_2$	4-class	65.5	65.0	-
TFN	2-class	75.2	-	0.760	H-DMS	5-class	60.4	60.2	0.594
LSTM(A)	2-class	73.5	-	0.703	UL-Fusion*	4-class	66.5	66.8	0.663
UL-Fusion*	2-class	72.5	72.5	0.730	DL-Fusion*	4-class	65.8	65.7	0.665
DL-Fusion*	2-class	71.8	71.8	0.720	Ours-HF	4-class	70.0	69.7	0.695
Ours-HF	2-class	74.1	74.4	0.744	Ours-VF	4-class	71.8	71.8	0.713
Ours-VF	2-class	75.3	75.3	0.755	Ours-FAF	4-class	**72.7**	**72.7**	**0.726**
Ours-FAF	2-class	**76.4**	**76.5**	**0.768**	Ours-FAF	5-class	**64.6**	**63.4**	**0.644**

Table 3.1: Comparison of models. *WA* = weighted accuracy. *UA* = unweighted accuracy. * denotes that we duplicated the method from cited research with the corresponding dataset in our experiment.

mentioning: (i) our model outperforms the baselines without using the low-level handcrafted acoustic features, indicating the sufficiency of MFSCs; (ii) the proposed approach achieves performance comparable to the model using text, audio, and visual data together [Zadeh et al., 2017]. This demonstrates that the visual features do not contribute as much during the fusion and prediction on MOSI; (iii) we notice that [Poria et al., 2017b] reports better accuracy (79.3%) on MOSI, but their model uses a set of utterances instead of a single utterance as input.

For emotion recognition, our model with FAF achieves 72.7% accuracy, outperforming all the baselines. The result shows the proposed model brings a significant accuracy gain to emotion recognition, demonstrating the pros of the fine-tuning attention structure. It also shows that word-level attention indeed helps extract emotional features. Compared to C-MKL$_2$ and SVM Trees that require feature selection before fusion and prediction, our model does not need an additional architecture to select features. We further evaluated our models on 5 emotion categories, including frustration. Our model shows 4.2% performance improvement over H-DMS and achieves 0.644 weighted-F1. As H-DMS only achieves 0.594 F1 and also uses low-level handcrafted features, our model is more robust and efficient.

Modality	MOSI		IEMOCAP	
	WA	F1	WA	F1
T	75.0	0.748	61.8	0.620
A	60.2	0.604	62.5	0.614
T+A	**76.4**	**0.768**	**72.7**	**0.726**

Table 3.2: Accuracy (%) and F1 score on text only (T), audio only (A), and multi-modality using FAF (T+A).

From Table 3.1, all the three proposed fusion strategies outperform UL-Fusion and DL-Fusion on both MOSI and IEMOCAP. Unlike utterance-level fusion that ignores the time-scale-sensitive associations across modalities, word-level fusion combines the modality-specific features for each word by aligning text and audio, allowing associative learning between the two modalities, similar to what humans do in natural conversation. The result indicates that the proposed methods improve the model performance by around 6% accuracy. We also notice that the structure with FAF outperforms the HF and VF on both MOSI and IEMOCAP dataset, which demonstrates the effectiveness and importance of the FAF strategy.

3.5.2 Modality and Generalization Analysis

From Table 3.2, we see that textual information dominates the sentiment prediction on MOSI and there is an only 1.4% accuracy improvement from fusing text and audio. However, on IEMOCAP, audio-only outperforms text-only, but as expected, there is a significant performance improvement by combining textual and audio. The difference in modality performance might because of the more significant role vocal delivery plays in emotional expression than in sentimental expression.

We further tested the generalizability of the proposed model. For sentiment generalization testing, we trained the model on MOSI and tested on the YouTube dataset

Approach	MOSI ↓ YouTube		IEMOCAP ↓ EmotiW	
	WA	F1	WA	F1
Ours-HF	62.9	0.627	59.3	0.584
Ours-VF	64.7	0.643	60.8	0.591
Ours-FAF	**66.2**	**0.665**	**61.4**	**0.608**

Table 3.3: Accuracy (%) and F1 score for generalization testing.

(Table 3.3), which achieves 66.2% accuracy and 0.665 F1 scores. For emotion recognition generalization testing, we tested the model (trained on IEMOCAP) on EmotiW and achieves 61.4% accuracy. The potential reasons that may influence the generalization are: (i) the biased labeling for different datasets (five annotators of MOSI vs one annotator of Youtube); (ii) incomplete utterance in YouTube dataset (such as "about", "he", etc.); (iii) without enough speech information (EmotiW is a wild audio-visual dataset that focuses on facial expression).

3.5.3 Visualize Attentions

Our model allows us to easily visualize the attention weights of text, audio, and fusion to better understand how the attention mechanism works. We introduce the emotional distribution visualizations for word-level acoustic attention (w_α_i), word-level textual attention (t_α_i), shared attention (s_α_i), and fine-tuning attention based on the FAF structure (u_α_i) for two example sentences (Figure 3.3). The color gradation represents the importance of the corresponding source data at the word-level.

Based on our visualization, the textual attention distribution (t_α_i) denotes the words that carry the most emotional significance, such as "hell" for *anger* (Figure 3.3 a). The textual attention shows that "don't", "like", and "west-sider" have similar weights in the *happy* example (Figure 3.3 b). It is hard to assign this sentence *happy*

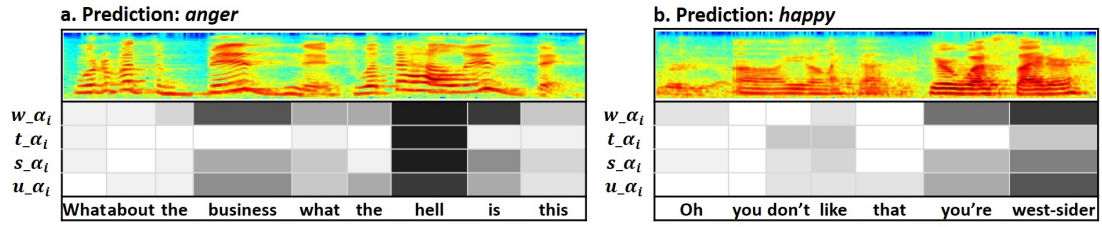


Figure 3.3: The overall system structure for multimodal hierarchical attention structure with word-level alignment.

given only the text attention. However, the acoustic attention focuses on "you're" and "west-sider", removing emphasis from "don't" and "like". The shared attention (s_α_i) and fine-tuning attention (u_α_i) successfully combine both textual and acoustic attentions and assign joint attention to the correct words, which demonstrates that the proposed method can capture emphasis from both modalities at the word-level.

3.6 Summary

There are several limitations and potential solutions worth mentioning: (i) the proposed architecture uses both the audio and text data to analyze the sentiments and emotions. However, not all the data sources contain or provide textual information. Many audio-visual emotion clips only have acoustic and visual information. The proposed architecture is more related to spoken language analysis than predicting the sentiments or emotions based on human speech. Automatic speech recognition provides a potential solution for generating the textual information from vocal signals. (ii) The word alignment can be easily applied to human speech. However, it is difficult to align the visual information with text, especially if the text only describes the video or audio. Incorporating visual information into an aligning model like ours would be an interesting research topic. (iii) The limited amount of multimodal sentiment analysis and emotion recognition data is a key issue for current research, especially for deep models

that require a large number of samples. Compared large unimodal sentiment analysis and emotion recognition datasets, the MOSI dataset only consists of 2199 sentence-level samples. In our experiments, the EmotiW and MOUD datasets could only be used for generalization analysis due to their small size. Larger and more general datasets are necessary for multimodal sentiment analysis and emotion recognition in the future.

In this paper, we proposed a deep multimodal architecture with hierarchical attention for sentiment and emotion classification. Our model aligned the text and audio at the word-level and applied attention distributions on textual word vectors, acoustic frame vectors, and acoustic word vectors. We introduced three fusion strategies with a CNN structure to combine word-level features to classify emotions. Our model outperforms the state-of-the-art methods and provides effective visualization of modality-specific features and fusion feature interpretation.

Chapter 4

Mutual Attentive Fusion Network

4.1 Introduction of Chapter

Even though the primary focus of previous research has been to classify utterance-level emotions based on a single data source (words, audio signal, facial expression, etc.), recent works demonstrate the necessity and benefits of multimodal architectures that combine heterogeneous inputs to predict emotion with joint modalities [Zadeh et al., 2017, Poria et al., 2015, Gu et al., 2018d]. Aside from multimodal analysis, more recent works employ dialogs and dyadic communication rather than single utterance as input to provide contextual information for emotion recognition [Poria et al., 2017b, Zadeh et al., 2018, Gu et al., 2018b]. In this paper, we focus on learning human emotional state based on dyadic verbal expressions. Specifically, we consider sequence and contextual information of verbal communication in the form of acoustic signals and linguistic content to predict utterance-level emotion [1].

Although previous approaches have achieved good performance, there still exist several challenges in multimodal dyadic emotion recognition: 1. Different sensor data require independent preprocessing and feature extraction designs due to the heterogeneous formats [Poria et al., 2015, Poria et al., 2016]. Using the appropriate approaches

[1] This work has been published in 2019 Proceedings of the 27th ACM International Conference on Multimedia [Gu et al., 2019].

to capture representative modality-specific features is critical to model performance. 2. The multiple modalities significantly increase the complexity of both the individual modalities and fusion model, especially for the recent deep learning-based architectures [Gu et al., 2018b, Chen et al., 2017]. To make an applicable and generalizable model for multimodal emotion recognition, it is necessary to consider the tradeoff between computational complexity and performance. 3. Little research provides solutions to uncertainty in label disagreement in emotion recognition. However, as emotion is an abstract and subjective concept, it is very common in both real-world scenarios and multimodal emotion datasets to have utterance-level data with diverse emotions from different people or annotators. Of the IEMOCAP dataset [Busso et al., 2008], 28.2% of utterance-level samples cannot be assigned to a specific emotion category due to disagreements from all annotators; only around 37.5% of utterance-level data have complete agreements. Most previous research uses only the completely-agreed data or applies majority vote on the labels [Zadeh et al., 2017, Poria et al., 2015, Gu et al., 2018d, Poria et al., 2017b, Zadeh et al., 2018, Gu et al., 2018b, Poria et al., 2016]. Unfortunately, these approaches abandon the disagreeing data and cannot fully reveal the actual emotional state. This restriction may cause a discontinuities or gaps during dyadic emotion recognition.

Addressing the issues above, we introduce a novel efficient dyadic fusion network that only relies on an attention mechanism to select informative features, combine unimodal features, and capture contextual information. We first design a sub-view attention based on the self-attention mechanism [Vaswani et al., 2017] for both the feature extraction models and fusion model. Unlike the previous approaches that use diverse and complex sub-embedding networks to extract modality-specific features [Gu et al., 2018b,

Poria et al., 2017c], we design two very simple but effective models with sub-view attention mechanisms to extract the textual and acoustic representations. We train the two independent modalities without considering contextual information during feature extraction. Our design allows fast convergence in a few training epochs. Then, we generate utterance-level acoustic and textual representations, respectively. To improve fusion efficiency, we introduce the sub-view attention layer to replace recurrent architectures in previous research [Poria et al., 2017b, Zadeh et al., 2018, Poria et al., 2017c]. We further facilitate attention-based modality fusion by introducing a mutual correlation attentive factor to learn the mutual attention distribution across different modalities. The learned acoustic or textual mutual representations are then fused with the original representations to finalize the information exchange in each sub-view attention layer. To solve the disagreeing annotation issue, for each utterance, we embed all concurrent labels into a k-dimensional vector (where k represents the number of classes) based on the label count and transform the categorical problem to a regression problem. This method allows the full use of each utterance and its label. We evaluate the proposed model on two published multimodal emotion recognition datasets. Our model significantly outperforms previous state-of-the-art research by 3.8%-7.5% accuracy, using a more efficient model. The main contribution of our paper can be summarized as:

- An efficient dyadic fusion network that mainly relies on an attention mechanism for feature extraction, modality fusion, and contextual representation learning.
- A novel mutual correlation attentive factor that automatically learns the associations across modalities in each sub-view attention layer to facilitate fusion.

- An effective solution and a detailed experimental analysis of the label disagreement issue that keeps sequence consistency and allows full use of labeled dialog data.

4.2 Related Work

A basic challenge for multimodal emotion recognition is to extract informative modality-specific features. Previous approaches can be separated into two categories: low-level hand-crafted features and abstract high-level representations. A large body of low-level features for both the text and audio branch-es has been proposed in previous decades, such as the bag of words and part-of-speech tagging for text, and the low-level descriptors with statistics for audio [Seppi et al., 2008, Savran et al., 2012, Eyben et al., 2010b, Degottex et al., 2014a]. However, the lack of high-level associations between features prevents improvements in the model performance. To overcome this issue, recent works used deep learning models to extract high-level representations from the low-level features, resulting in performance improvements. A convolutional neural network was used to extract the textual features from the embedding word vectors in [Poria et al., 2015, Poria et al., 2017b]. The long short-term memory network was applied to both the text and audio branch to capture the temporal features [Poria et al., 2017b, Rajagopalan et al., 2016, Liang et al., 2018a]. More recently, attention mechanisms were integrated with recurrent neural networks to select informative textual and acoustic features [Zadeh et al., 2018, Gu et al., 2018b, Poria et al., 2017c]. Compared to the manually handcrafted features, the deep models allow automatic feature extraction and can learn representative associations from low-level features. Later, word-level feature extraction was introduced [Gu et al., 2018d, Chen et al., 2017] to

further improve modality-specific feature extraction. Most previous works focused on using single utterance to identity emotion [Poria et al., 2015, Gu et al., 2018d], while the more recent works started combining the surrounding utterances as context to provide extra information for utterance-level emotion recognition [Poria et al., 2017b, Zadeh et al., 2018, Gu et al., 2018b]. These approaches require the ability to extract modality-specific features not only from a single utterance, but also from the surrounding utterances. Hence, designing an effective and efficient structure to select the informative contextual features is necessary in multimodal emotion recognition.

In addition, modality fusion is challenging due to the heterogeneous inputs. Early research applied late fusion to combine prediction results by some algebraic rules [Wöllmer et al., 2013b], avoiding the difficulty of combining heterogeneous features. However, such approaches ignore associations across modalities and fail to measure mutual correlations. To address the above issue, recent works proposed deep fusion networks to combine modality-specific representations at the feature-level [Zadeh et al., 2017, Poria et al., 2016, Liu et al., 2018], which allows significant performance improvement. To further measure the temporal and context information, a multi-attention recurrent network was proposed [Zadeh et al., 2018] to learn both modality-specific and cross-view interactions over time. A local-global ranking fusion strategy integrated with LSTM and a recurrent multistage fusion model were introduced [Liang et al., 2018b, Liang et al., 2018a] to fuse the features in a timeline. A hierarchical encoder-decoder structure was proposed, which relied on an LSTM to encode modality-specific features and decode the prediction in sequence. A context-dependent model using two unidirectional LSTMs to predict human emotion from context utterances was proposed in

[Gu et al., 2018b]. Although most used recurrent neural networks to identify temporal or context information during emotion recognition, we argue that this is neither necessary nor efficient because: 1. A specific word or utterance may directly indicate the emotional state and then dominate the final decision. Instead of word-by-word or utterance-by-utterance feature extraction in RNNs, learning the informative word or utterance representations is more helpful. 2. The RNNs require more training time compared with other approaches because they can only compute sequentially.

To address the above issues, we propose a dyadic fusion network that mainly relies on attention mechanisms to extract contextual features and fuse the multimodal information.

4.3 Methodology

4.3.1 System Overview

Our model consists of three major modules: modality-specific feature extraction, modality fusion, and decision making. To facilitate the fusion of heterogeneous inputs, we first introduce the sub-view attention structure and extract modality-specific features for each single utterance. Then, we treat the surrounding utterances as the context of the current utterance and concatenate the generated utterance representations in sequential order as the input for modality fusion. Specifically, for the current utterance, we consider all the previous utterances in the same dialog as the context information. We further design a mutual correlation attentive factor (MCAF) combined with sub-view attention structure to fuse the contextual modality-specific representations. We use a four-layer sub-view attention with MCAF to select the features, learn cross-modality associations, and compute the attention distribution over the entire dialog or dyadic

sequence for each modality. Finally, we concatenate the two generated fusion representations and introduce 1D average-pooling to generate the final joint representation. The model is trained with a regression strategy to predict utterance-level emotion.

4.3.2 Sub-View Attention Mechanism

The sub-view attention mechanism is the foundation of both feature extraction and modality fusion. Inspired by the work in [Vaswani et al., 2017] that proposes a multi-head self-attention mechanism in machine translation, we replace recurrent approaches with attention for emotion recognition because: 1. The temporal features are not the most critical information for emotion detection on both utterance-level and dialog-level data. Most dyadic communication and verbal utterances are short sentences, so a specific word or utterance may directly indicate the emotional state and dominate the final decision. Unlike RNNs that learn features word-by-word or utterance-by-utterance, attention directly computes the importance score of each word or utterance, providing an intuitive weighted representation to help the final decision. 2. Because the attention computation can be processed in parallel (rather than sequentially, as in recurrent approaches), attention architectures are more efficient in both training and inference [Vaswani et al., 2017]. This significantly reduces the model size and computational complexity, especially for multimodal research.

The basic concept of self-attention can be understood as a weighted computation of each value using the corresponding overall mapping of query-key sets (shown in Fig.4.1). As suggested in self-attention [10], we first generate the query (q), key (k), and value (v) by computing the linear projection of the input i with different parameter matrices

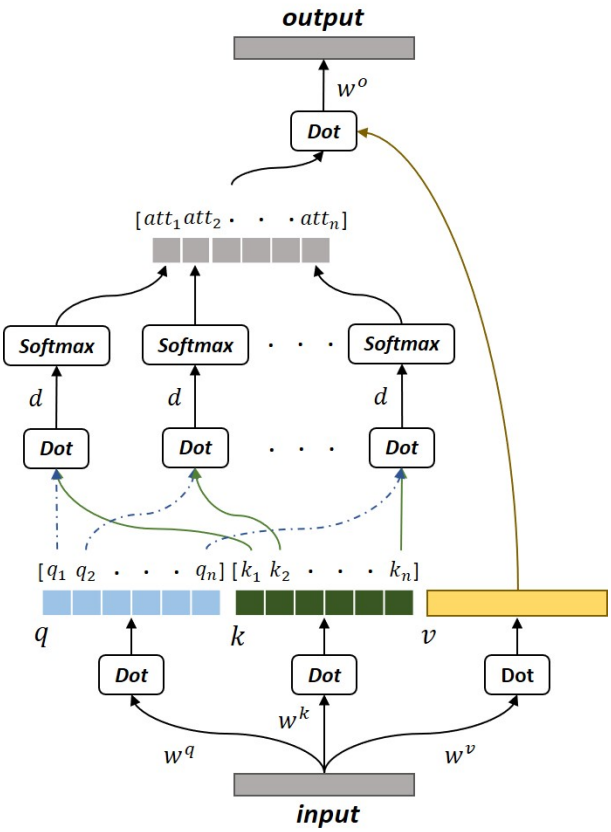


Figure 4.1: Sub-view attention mechanism

(w^q,w^k,and w^v), respectively:

$$q, k, v = linear(iw^q, iw^k, iw^v) \tag{4.1}$$

Instead of applying multiple linear operations with different learnable projection parameters to generate multiple q, k, and v as in multi-head self-attention, we only compute a single linear projection for q, k, and v, repectively. Then, we separate the q and k into n sub-vectors to further compute the attention over the individual q_j and k_j:

$$att_j = softmax(\frac{q_j k_j^T}{\sqrt{d}}), j \in [1, n] \tag{4.2}$$

where d is the scale dimension and $n*d$ equals the input dimension (i). The generated

att_j can be intuitively seen as the sub-view attention based on the jth query-key pair. The final output o can be represented as:

$$o = [concate(att_1, att_2, \ldots, att_n)v]w^o \tag{4.3}$$

where the w^o is the parameter matrix of the output linear projection. The proposed sub-view attention focuses on learning the attention distribution over the sub-space of each query-key pair. Because we only process a single linear operation rather than generate multiple sub-projected queries, keys, and values, the model further reduces the computational cost and improves model efficiency.

4.3.3 Modality-specific Feature Extraction

We first train the textual and acoustic modalities independently to generate the utterance-level modality-specific representations. Because our work focuses on learning the dialog-level emotional state from multiple utterance-level representations, an effective and efficient architecture is necessary for model generalization. Unlike the structures in [Poria et al., 2017b, Gu et al., 2018b] that consist of diverse models and multiple deep networks to extract modality-specific features, we design two effective shallow neural networks to extract unimodal features. We leave the contextual information learning for the modality fusion stage and train the unimodalities without using the surrounding utterances. This means each representation only relies on the current item in the verbal transcript or audio stream.

To extract the textual representations for each utterance, as shown in Fig.4.2, we first embed each word into a 200-dimensional vector using pretrained word vectors from *Glove* [Pennington et al., 2014]. Then, we feed the embedded word vectors into

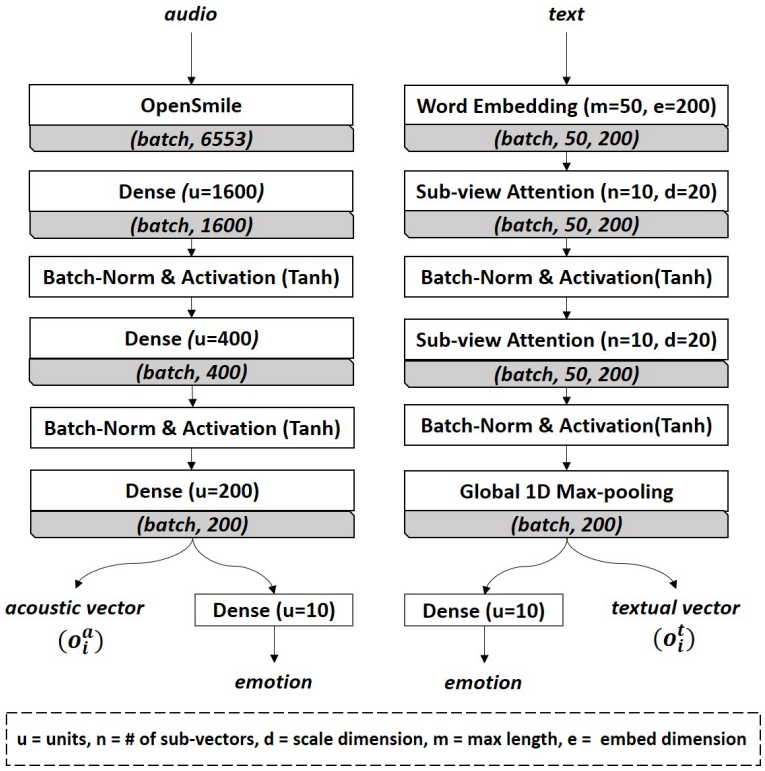


Figure 4.2: Modality-specific feature extraction

the sub-view attention layer to compute the attentive dependencies and generate the weighted representation for each word. The output of the layer has the same dimension as the input; we set two sub-view attention layers to learn the features. The output from the last attention layer directly connects to a global 1D max-pooling operation to form the utterance-level textual representation. The final output is a 200-dimensinal feature vector.

To generate the acoustic representations, we directly use the *openSmile* toolkit [Eyben et al., 2010b] to extract low-level descriptors (LLDs) for each utterance-level audio stream to reduce the model complexity. The feature set contains 6553 features including voice intensity, pitch, MFCCs, etc. We apply three dense layers to learn the high-level associations from the LLDs and reduce the dimension of the acoustic representation. As shown in Fig.4.2, the acoustic representation for each utterance is also a

200-dimensional vector.

We format the output utterance-level representations into the dialog-level based on the sequence order. Each input sample of the fusion module becomes a 2D matrix with [h, 200] as the shape. The h indicates the number of all utterances from the first utterance in the dialog to the current utterance. We perform zero-padding to align all samples based on the longest dialog from the dataset.

4.3.4 Modality Fusion with Mutual Correlation Attentive Factor

Instead of feeding dialog-level samples into a recurrent neural network in sequential order as most previous research did [Poria et al., 2017b, Zadeh et al., 2018], we design a mutual correlation attentive factor integrated with the proposed sub-view attention to extract the dialog-level features and learn cross-modality associations simultaneously. As shown in Fig.4.3, the fusion model first applies the same sub-view attention structure to learn the dialog-level attentive dependencies on the textual and acoustic representations. Unlike the original sub-view attention that simply relies on the independent textual k^t or acoustic k^a to compute the attention, we introduce two learnable factors l^t and l^a to fuse the keys for each branch, respectively:

$$k^{t^*} = k^t + l^t k^a \tag{4.4}$$

$$k^{a^*} = k^a + l^a k^t \tag{4.5}$$

The fused textual key (k^{t^*}) and acoustic key (k^{a^*}) continue to separately compute the textual and acoustic attention using equation (2). The two factors learn the mutual correlations between independent keys, helping the model compute attention over both

the textual and acoustic branches. This allows model fusion inside each attention layer. Fig.4.3 shows the details of the mutual correlation attentive factor.

We set four sub-view attention layers with mutual correlation attentive factors to compute attentions on each utterance and fuse the textual and acoustic modalities. The output weighted vectors contain the attentions of both the modality-specific and cross-modality context; the final outputs are o^t for textual representation and o^a for acoustic representation.

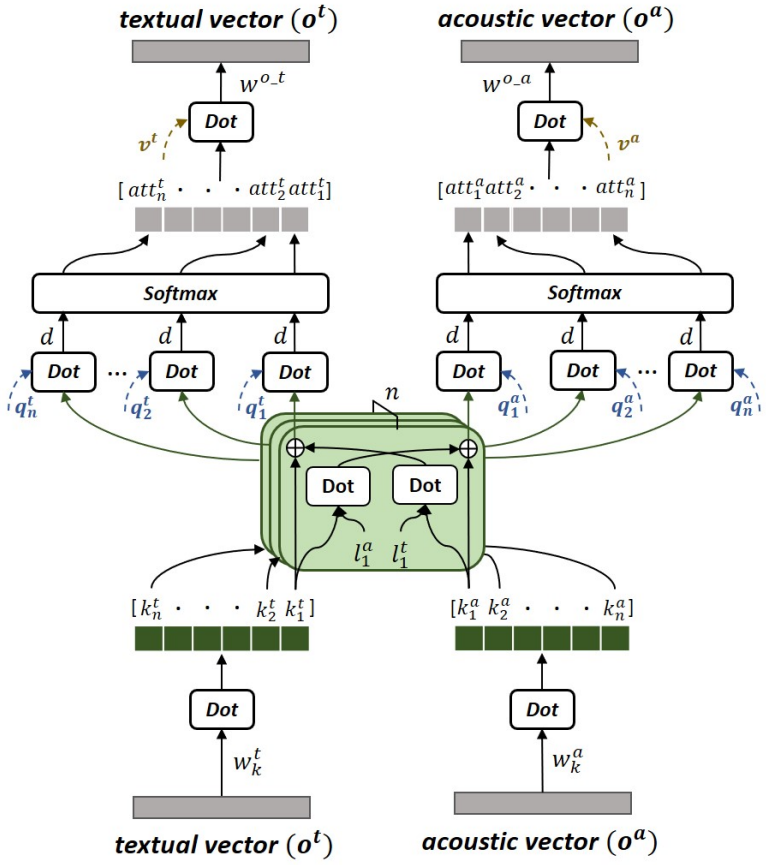


Figure 4.3: Mutual correlation attentive factors (MCAF) in sub-view attention for modality fusion

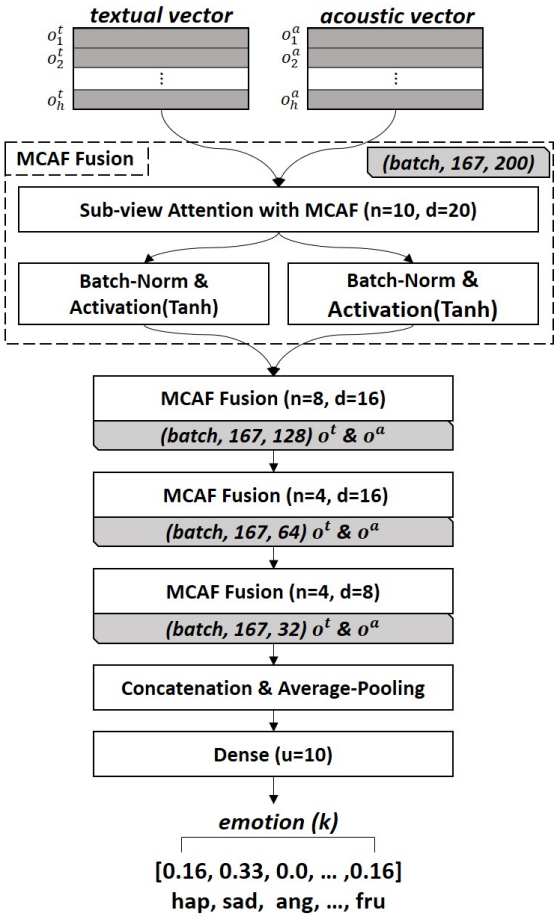


Figure 4.4: Dyadic fusion network

4.3.5 Decision Making

Fig.4.4 shows the overall structure of the dyadic fusion network. As suggested by the self-attention mechanism [Vaswani et al., 2017], we first connect each MCAF sub-view attention layer with a batch normalization layer [Ioffe and Szegedy, 2015] and an activation function. The proposed attention mechanism learns the attention on utterances over the entire dialog, so each utterance has already been represented by the weighted score to indicate the corresponding importance in dialog. We do not use RNNs to generate the contextual vectors because the attention mechanism allows each utterance to learn the dependencies from other utterances. Since each utterance

has already integrated the information from all other utterances, there is no need for the model to learn the temporal information step by step. Removing the recurrent neural networks also increases the training speed due to the parallel computation of attention. To make the final decision, we concatenate the generated o^t and o^a to form the joint representation and use an average pooling and dense layer to form the final representation (shown in Fig.4.4).

Compared to the previous approaches that classify emotion only based on all-agreeing or majority-voted labels [Zadeh et al., 2017, Poria et al., 2015, Gu et al., 2018d, Poria et al., 2017b], we embed the labels from all annotators into a k dimensional vector based on the number of classes and scale the vector to sum to one. We fit the final representation with the scaled labels in a regression method because: 1. The scaled vectors reveal the actual emotional state and allow the full use of the entire dataset. Some previous works assign the disagreeing labels to the '*Other*' category during modeling [Gu et al., 2018b], which is inappropriate because the placeholder category may consist of contradicting emotions. For example, '*I just don't. It's stupid.*' (with the labels *Anger*, *Disgust*, *Frustration*) and '*I've been ready a long, long time.*' (with the labels *Excited*, *Happiness*, *Surprise*) were both assigned to 'other' due to disagreeing labels, although they contain opposite emotional states. 2. The regression approach trains the model to output a mixed ratio, which has been demonstrated effective in [Tokozume et al., 2017]. We finally compute the argmax based on the output to transform the regression metric into a categorical metric.

4.4 Experiments

4.4.1 Dataset Configuration

We evaluated our model on two published multimodal emotion recognition datasets: IEMOCAP and MELD.

IEMOCAP: The Interactive Emotional Dyadic Motion Capture database is an acted, multimodal, multi-speaker emotion recognition dataset recorded across 5 sessions including 12 hours of video, speech, and text [Busso et al., 2008]. For this study, we only use audio and text data. The dataset consists of 10039 utterances from 151 dialogs and contains 10 categories including 'neutral', 'exciting', 'sadness', 'frustration', 'happiness', 'angry', 'other', 'surprised', 'disgust', and 'fear'. For each utterance, we include the labels from all annotators and embed it as a 10-dimensional vector. We follow previous research to split the data into training, validation, and testing sets at the session level [4, 5]. The split considers the speakers independent. The final dataset has 3 sessions for training, 1 session for validation, and 1 session for testing.

MELD: Multimodal EmotionLines Dataset (MELD) is a multimodal and multi-speaker dataset that enhances and extends EmotionLines [Poria et al., 2018, Chen et al., 2018]. It contains about 1400 dialogues and 13000 utterances with video, speech, and text from the Friends TV series. Its seven emotions include 'anger', 'disgust', 'sadness', 'joy', 'neutral', 'surprise' and 'fear'. The dataset has already been split into training (1039 dialogues with 9989 utterances), testing (114 dialogues with 1109 utterances), and dev (280 dialogues with 2610 utterances) data.

4.4.2 Baselines

We compare the performance of our model to the following baselines for the multimodal emotion recognition task.

SVM: an SVM classifier trained on the concatenation of text and audio features [Rosas et al., 2013].

RF: a random forest model that also uses the concatenated text and audio branch features [Breiman, 2001].

C-MKL: a convolutional neural network with a multiple kernel learning strategy to predict emotion and sentiment based on multimodal data [Poria et al., 2016].

EF-LSTM: an early fusion strategy to concatenate the inputs from different modalities at each time step and apply a single LSTM to learn temporal information from the joint representations [Zadeh et al., 2018].

BC-LSTM: a context-dependent model using two unidirectional LSTMs to predict human sentiment and emotion, which can identify information from context utterances [Poria et al., 2017b].

MV-LSTM: a recurrent model to capture both modality-specific and cross-view interactions over time or structured outputs from multiple modalities [Rajagopalan et al., 2016].

TFN: a tensor fusion network that uses a multi-dimensional tensor to learn view-specific and cross-view dynamics across three modalities for emotion recognition and sentiment analysis tasks [Zadeh et al., 2017].

HAW: a multimodal structure using hierarchical attention with word-level alignment to utterance-level sentiment and emotion [Gu et al., 2018d].

AMN: an attentive multimodal network using a hierarchical encoder-decoder to

predict the sentiment and emotions with contextual information [Gu et al., 2018b].

MARN: a multi-attention recurrent network that explicitly models both view-specific and cross-view dynamics in the network through time by using a specific neural component called Multi-attention Block (MAB) [Zadeh et al., 2018].

4.4.3 Implementation

We implement the model with *Keras* [Chollet et al., 2015] and *Tensorflow* [Abadi et al., 2016] backend. We use normalized low-level features extracted by *OpenSmile* based on each feature type with zero mean and unit variance. The detailed information of each layer is shown in Fig.4.2 and Fig.4.4. The modality feature extraction module and modality fusion module are trained on the same training-validation-testing split. We set the learning rate to 0.0001 and use the Adam optimizer with mean square error loss for both the pretraining and fusion modeling. We compute the argmax of the output from our model to indicate the prediction class. To make a fair comparison with previous research, we reimplement the baseline models from the source code provided by the authors using our dataset splits. For the models that cannot be applied on two modalities (TFN) or that do not have source code (EF-LSTM), we directly use the performance reported in [Zadeh et al., 2018]. All the models are trained on the entire dataset, rather than on majority-voted or all-agreement data as in previous research. As suggested in [Gu et al., 2018b], We assign the disagreeing labels to the *'Other'* category for all baselines.

	Modality	Context	Acc.(%)	F1-Score
SVM	T+A	no	27.2(↑ 24.4)	27.3(↑ 23.0)
RF	T+A	no	30.5(↑ 21.1)	22.1(↑ 28.2)
C-MKL	T+A	no	37.0(↑ 14.6)	36.1(↑ 14.2)
EF-LSTM	T+A+V	no	34.1(↑ 17.5)	32.3(↑ 18.0)
BC-LSTM	T+A	yes	38.9(↑ 12.7)	38.1(↑ 12.2)
MV-LSTM	T+A	yes	37.2(↑ 14.4)	37.2(↑ 13.1)
TFN	T+A+V	no	36.0(↑ 14.6)	34.5(↑ 15.8)
HAW	T+A	no	40.8(↑ 10.8)	40.8(↑ 9.5)
AMN	T+A	yes	43.4(↑ 8.2)	43.3(↑ 7.0)
MARN	T+A	yes	44.1(↑ 7.5)	43.9(↑ 6.4)
Ours(cate)	T+A	yes	47.3(↑ 4.3)	47.2(↑ 3.1)
Ours(reg)	T+A	yes	**51.6**	**50.3**

Table 4.1: Emotion recognition result on IEMOCAP dataset. Following previous research, the metric computation based on 9 categories (without *'other'*).

4.5 Result Analysis

4.5.1 Comparison with Baselines

We first compare our model with the baselines and the state-of-the-art on IEMOCAP. Following previous research [Zadeh et al., 2018], we compare the model performance without considering the *'Other'* category. The result shown in Table 1 indicates that the performance of our model significantly outperforms previous approaches at both accuracy and weighted F1-score. The proposed dyadic fusion network using mutual correlation attentive factors gains 7.5% accuracy and 6.4 F1-score improvement over the previous state-of-the-art. We have the following findings from Table 4.1: 1. The significant performance improvement shows that the proposed architecture is effective for multi-class classification. Even without using the visual features, our model still achieves the best performance on IEMOCAP. 2. Using contextual information indeed helps emotion recognition. The structures that consider previous utterances during prediction perform better than the models that only rely on a single utterance; this demonstrates the necessity of context.

	Ang		Joy		Neu		Sad		Sur	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN(T)	75.4	74.8	65.1	64.7	63.1	62.7	77.4	72.1	75.5	71.3
BiLSTM(T)	74.8	73.7	65.0	64.3	63.3	63.1	77.6	72.2	74.7	71.2
BiLSTM(T)	68.3	64.2	63.2	60.1	56.5	54.3	69.3	62.5	69.4	65.9
BiLSTM(T+A)	75.9	74.1	67.4	66.3	**65.8**	64.7	80.2	74.4	76.1	73.6
Ours(T+A)	**79.4**	**75.3**	**70.4**	**70.1**	65.7	**65.4**	**84.0**	**79.2**	**78.3**	**74.0**

Table 4.2: Emotion recognition result on MELD dataset (%). the metric computation based on binary classification for each emotion. Ang=anger, Neu=neutral, Sur=surprise.

We also evaluate the model performance on MELD. Because it is a newly released dataset, there are very few works using MELD. We directly compared our model with the baseline models proposed in [Poria et al., 2018]. Due to the imbalanced emotion split in MELD, we conduct binary classification during experiments. The result in Table 4.2 shows that our model outperforms the baselines on both accuracy and F1-score in anger, joy, sad, and surprise. We notice that our model only achieves 65.7% in the neutral class, but all baselines have relatively bad performance there. After analyzing the raw data, we found a significant number of neutral emotion samples with only very subtle differences compared to the other emotions. We believe the ambiguity of neural samples extremely reduces the performance of neural detection. Since the MELD dataset only has one annotator for each utterance, we argue that the data may have personal bias and some inaccurate emotion labels.

4.5.2 Quantitative Analysis

We further evaluate our model by comparing the performance of unimodal and multimodal structures (shown in Table 4.3). We compute individual accuracy (9 category) and list four general emotions including Ang (*'angry'*), Neu (*'neutral'* + *'frustration'*), Sad (*'sadness'*), and Hap (*'happiness'* + *'exciting'*). The result indicates that the

	Acc.	Weighted-F1	Ang	Neu	Sad	Hap
Ours(T)	43.8	44.3	33.4	50.6	45.5	36.6
Ours(A)	36.7	36.7	31.8	44.7	38.1	24.1
w/o-Context	47.2	46.2	**35.8**	48.1	67.2	48.8
w/o-MCAF	46.7	44.5	18.4	52.7	66.2	47.8
Ours(T+A)	**51.6**	**60.3**	31.8	**54.1**	**74.1**	**61.0**

Table 4.3: Quantitative analysis on IEMOCAP dataset (%). Ang = anger, Neu = neutral+frustration, Hap = happy+exciting.

textual modality performs better than acoustic modality in general. The multimodal structure significantly improves the performance on Neu, Sad, and Hap. Even with a slight performance decrease on Ang, combining two modalities still provides 7.8% accuracy improvement from textual modality and 14.9% accuracy improvement from acoustic modality. This demonstrates the helpfulness of applying multimodal structure. In addition, the proposed unimodal structures achieve comparable performance to the baseline multimodal structures, especially for the text modality, which achieves 43.8% accuracy. This indicates the proposed modality-specific models and the regression training strategy are more effective than previous approaches.

We design an experiment on our model without using contextual information. The only difference between the with- and without-context model is that the model without context only uses a single utterance representation as the fusion input and we set zero values as the context information. As shown in Table 4.3, using contextual information improves 4.4% accuracy and 4.1 F1-score, which shows that context contains additional information that can facilitate emotion recognition. The model without context performs better than the contextual model on Ang, which means context information does not provide positive contribution during the final prediction in our experiment.

To illustrate the performance of the proposed mutual correlation attentive factors, we compare the model with and without MCAF. The result shows that using MCAF

	Trainable Parameters	Training FLOPs	Training Speed (ms/per epoch)	Acc. (%)
Ours(T)	2.9×10^7	5.7×10^7	2.3×10^7	38.9
Ours	1.3×10^7	2.7×10^7	1.2×10^4	51.6

Table 4.4: Comparison of training cost on IEMOCAP dataset.

increases 4.9% accuracy and 5.8 F1-score on the IEMOCAP dataset. The model without MCAF only achieves 18.4% accuracy on Ang and the MCAF improves the performance by 18.5% accuracy. The better performance on both the overall and specific emotion categories demonstrates the usefulness of the proposed mutual correlation attentive factor.

We also compute the training cost of our model using three metrics: trainable parameters, number of floating-point operations, and the average training speed per epoch. We compare our model with the BC-LSTM approach that also considers contextual information during modeling. To make a fair comparison, we reimplemented their approach with the same train/dev/test set (without using visual data) and we trained both models on an NVIDIA GTX 1018ti with the same framework environment. Table 4.4 shows the training cost of the entire architecture including both the feature extraction and modality fusion. The result indicates that the proposed approach significantly reduces the training costs on all three metrics. Our model outperforms the BC-LSTM approach by 12.7% accuracy but only requires about half training cost, demonstrating the efficiency of the proposed network.

4.5.3 Disagreeing Annotation Analysis

Since disagreeing annotations are very common in most emotion datasets that consist of multiple annotators, giving an appropriate solution and a detailed analysis

for the disagreeing data is helpful and necessary for model generalization. Unfortunately, most previous approaches simply remove these samples in modeling and very rarely contain detailed analysis [Zadeh et al., 2017, Poria et al., 2015, Gu et al., 2018d, Poria et al., 2017b, Zadeh et al., 2018, Gu et al., 2018b, Poria et al., 2016]. In this section, we provide an analysis of the samples that cannot be assigned to a category in IEMOCAP.

As shown in Table 4.5, around 25% of utterances have disagreeing annotations in all three sets. Simply abandoning this data may cause incomplete dialogue. This gap may further influence contextual feature extraction and the prediction accuracy of the emotion state change in dyadic communication. Unlike all previous works, the proposed regression approach allows our model to fully use all data and simultaneously keep emotional information from the disagreeing labels, which maintains the consistency of the data. To analyze the disagreeing annotations, we first treat the disagreeing labels as multi-labels and compute the average precision for each category. As shown in Table 4.5, *'exciting'* and *'anger'* achieve 86.8 and 83.8 average precision, and the mean average precision of the overall multi-label samples is 52.0; this demonstrates our model can successfully learn multiple emotions and reveal actual emotional state for disagreeing data. We further compare the performance of the proposed regression approach and the categorical approach (directly assigning all disagreeing labels into *'other'* category). The result in Table 1 shows the regression approach increases 4.3% accuracy and 3.1 F1-score, showing that using disagreeing annotation data with regression training can provide extra information to improve emotion recognition.

Session Split	Train Set	Dev Set	Test Set
3/1/1	1405/5800	572/2136	564/2103

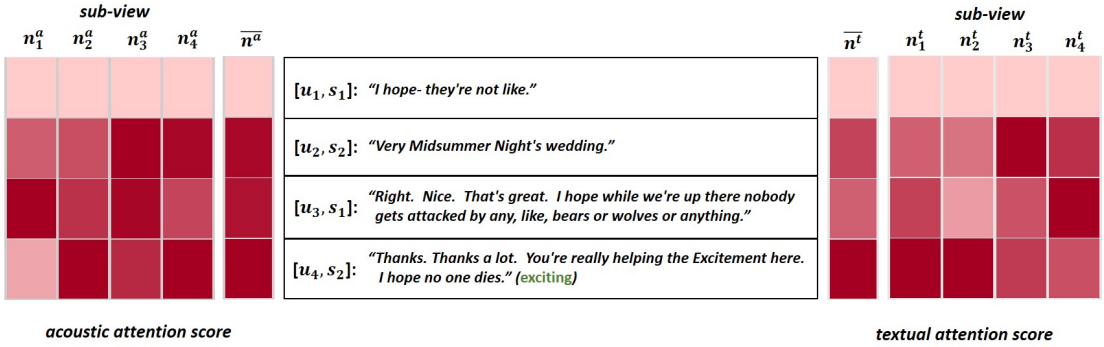
	neu	exc	sad	fru	hap	ang	oth	sur	dis	fea
P/A	12.8	18.5	11.1	15.8	16.0	17.9	3.40	3.50	0.50	0.00
AP	60.3	86.8	52.0	74.1	75.2	83.8	16.2	16.7	2.60	0.00

Table 4.5: Analysis of disagreement annotation on IEMOCAP dataset. number of disagreement annotation utterance/total utterances. P/A = number of the positive samples / number of total samples. AP = average precision.

4.5.4 Attention Visualization

In this section, we provide an example of the sub-view attention in Fig 4.5 to help human interpretation of the model. We plot the attention score (att_j, in equation (2)) of both textual and acoustic branches from the last MCAF fusion layer, respectively. The color gradation indicates the importance of the current utterance over the last utterance. In the example, the model predicts the emotion of the last utterance (u_4) based on both u_4 and the previous three utterances, which can be seen as contextual information for the u_4. Each branch consists of four sub-view attention scores. To facilitate understanding of the visualization, we compute the average scores of the four sub-scores to epresent the importance of the current utterance. As shown in Fig 4.5, the textual branch focuses on the last utterance itself and pays less attention to the first utterance. Our attention mechanism successfully measures the change of emotional state from 'neutral' to *'exciting'* in this example, which helps the model assign the last sentence to the correct category. For the acoustic branch, the last three utterances almost equally contribute to the final prediction. Both the textual and acoustic branches have already shared attention with each other due to the mutual correlation attention factor. This means the textual attention scores were decided not only by the textual representation, but also by the acoustic representation (similarly, for acoustic attention

scores). The visualization of the textual and acoustic attention scores can be intuitively understood as joint attention scores for each branch, respectively



$\overline{n^t}$, $\overline{n^a}$: average sub-view scores. n_i^t, n_i^a: sub-view attention score. u_i: the index of the utterances. s_1, s_2: speaker IDs.

Figure 4.5: Attention visualization.

4.6 Summary

In this paper, we introduced a dyadic fusion network that mainly relies on attention to extract contextual features and fuse multimodal information. We first used two effective light-weight modality-specific feature extractors to generate non-contextual representations for each utterance. Then, we combine the surrounding utterance representations as contextual input for modality fusion network. We designed a mutual correlation attentive factor integrated with the proposed sub-view attention mechanism to select representative vectors and learn cross-modal associations. We generated the labels for each utterance by embedding the corresponding labels from all annotators as a vector and used a regression approach to make the final decision. To the best of our knowledge, our work is the first one to provide a detailed analysis and solution on the disagreeing label issue. The experimental results show that our model significantly outperforms the previous approaches with less training cost. The results demonstrate the effectiveness

and efficiency of the proposed sub-view attention, mutual correlation attentive factor, and regression modeling strategy. Finally, we give a visualization of the attention to help human interpretation.

Chapter 5

Human Conversation Analysis Using Textual, Acoustic, And Visual Inputs

5.1 Introduction of Chapter

Human conversation analysis [1], including emotion recognition, sentiment analysis and speaker trait detection, is useful in many real-world applications such as medical support, activity recognition, chatbots, etc. Aside from challenges in sensor engineering and speech recognition, conversation understanding is still difficult because: (1) *Meaning can be expressed through different media.* A positive attitude can be expressed by words, facial expressions, and intonation, which are often captured by different sensors, requiring a feature fusion mechanism. Furthermore, we shall consider strategies that synchronize the input at the word-level because word is an important basic unit of meaning. (2) *Different sensors may indicate contradicting meanings.* For example, one person can pretend to be happy by saying happy words but with a sad face. Simply merging the features extracted from different modalities may confuse the system. The correct prediction can only be made by selecting a representative input modality and observing the context. (3) *The emotion, sentiment and traits during conversation may or may not change over time.* Most traditional conversation understanding strategies

[1] This work has been published in 2018 Proceedings of the 26th ACM International Conference on Multimedia [**?**].

only make a single prediction per conversation, which is inadequate for real-world applications. Furthermore, as conversations may contain multiple attributes (such as both happy and exciting), we need a flexible model that is able to perform classification, regression, and multi-label classification with only a slight modification.

To synchronize different sensor inputs, we introduce a feature extraction strategy that first aligns the raw text, audio, and video at the word level. Unlike sentence-level feature extraction and synchronization [Zadeh et al., 2016, Poria et al., 2015, Poria et al., 2016], our word-level feature extraction breaks down the features to a finer granularity with more details. Unlike direct feature fusion [Poria et al., 2016, Gu et al., 2018a], we proposed a fusion strategy with learned modality attention. The modality attention first identifies the importance of each input modality, then extends the importance of each modality to each feature dimension within that modality. Finally, to build a system that is both accurate and flexible, we designed a hierarchical encoder-decoder structure. The hierarchical structure first encodes the multimodal data into word-level features. The conversation-level encoder further selects important information from word-level features with temporal attention and represents all the conversation-level features as a vector. Considering that emotion and sentiment may change over a conversation and that multiple traits may be present simultaneously, our hierarchical decoder structure first decodes features at each time instance. Then, the attribute decoder will further decode the feature vector at each time instance into attributes at that time.

We tested our model on five published datasets. Our model outperformed the most recent state-of-the-art systems: emotion recognition with IEMOCAP [Busso et al., 2013] on classification; sentiment analysis with MOSI [Zadeh et al., 2016] on both classification and regression; and trait analysis with POM [Park et al., 2014] on classification. In

addition, our hierarchical encoder-decoder system was able to make multi-label predictions (predict multiple traits at once) and achieved performance comparable with most recent research which used 11 individual models. We further tested the generalizability of our system by training it on IEMOCAP and MOSI and testing on EmotiW and MOUD, respectively. Our system outperformed most recent state-of-the-art systems [Poria et al., 2017b] on the same transfer learning task by 6.8% accuracy. We further visualized our modality attention mechanism for modality fusion and temporal attention for encoding, demonstrating that the introduced modality attention model is able to select representative input modalities for sensor fusion. Our contributions include:

1. A word-level feature extraction strategy that is able to synchronize and extract features from different input modalities at the word level

2. A sensor fusion strategy with modality attention that can identify the importance of each modality and the importance of features within each modality.

3. A hierarchical encoder-decoder framework. The encoder encodes features from low level (word level) to high level (conversation level). The decoder first decodes the abstract features into attribute profile at each time instance, and then decodes the attribute profile at each time instance to individual attributes.

5.2 Related Work

Research on multimodal conversation understanding can be divided into three generations. The first generation used low-level handcrafted features for different modalities, including lexical representations for text, low-level descriptors (LLDs) for audio, and facial characteristic points (FCPs) for video [Rosas et al., 2013, Rozgic et al., 2012,

Savran et al., 2012]. Instead of simply concatenating the extracted modality-specific features as the final feature representation [Rosas et al., 2013], different shallow fusion strategies were introduced to learn the associations across different modalities, including Bayesian filtering [Savran et al., 2012] and ensemble SVM trees [Rozgic et al., 2012]. Handcrafted features, however, do not generalize well to different application scenarios, and modality-specific fusion strategies cannot effectively model the complex correlations between spatial and temporal information.

The second-generation systems tried address the issues caused by manually crafted features by extracting high-level features using deep learning. Convolutional neural networks (CNNs) were used for visual feature extraction [Krizhevsky et al., 2012] and recurrent neural networks (RNNs) (i.e. gated recurrent units (GRUs) and long short-term memories (LSTMs) [Poria et al., 2015, Poria et al., 2016, Gu et al., 2018d]) were used for audio and text feature extraction. RNNs were also used for learning long-term dependencies for the fusion of audio and video data [Wöllmer et al., 2013b]. Deep feature extractors can automatically learn features that are general and representative compared with manually-crafted features. Combining features learned from different input modalities with different time scales, however, remains a challenge [Poria et al., 2015, Poria et al., 2016]. Researchers tried to address modality fusion by applying decision-level fusion such as voting [Wöllmer et al., 2013b], but this approach failed to learn correlations between features extracted from different modalities over time.

Recent systems focus on model-level fusion, generating a shared representation of different modalities [Poria et al., 2017b, Chen et al., 2017, Zadeh et al., 2017, Zadeh et al., 2018]. Previous work directly used gated multimodal embedded fusion structures to fuse the

raw features, ignoring the temporal associations of individual modalities [Chen et al., 2017]. To avoid this problem, our system uses individual LSTMs to extract modality-specific features and learns their local temporal associations before fusion. Some researchers used only CNN and openSMILE to extract features [Poria et al., 2017b, Eyben et al., 2010a], but we introduce an attention-based LSTM structure to select informative word-level features. Instead of directly combining sentence-level features and ignoring the temporal associations [Zadeh et al., 2017], our hierarchical encoder-decoder LSTM learns the temporal associations at the word-level and conversation-level. We also applied temporal attention to select informative shared word-level representations to further improve the system's performance. The most recent research introduced a multi-attention recurrent network to fuse the modality features and learn the temporal associations [Zadeh et al., 2018], which achieved state-of-the-art performance on published multimodal datasets. However, this structure cannot learn the correlations across different modalities. Different modalities may have different importance during feature fusion; for example, acoustic features play a more important role in emotion recognition, but less in sentiment analysis. To model this information, we designed a modality fusion strategy that dynamically assigns the importance weights for input modalities.

5.3 Attentive Multimodal Networks with Hierarchical Encoder-Decoder

5.3.1 System Overview

We designed our system to be generalizable and flexible. To achieve generalizability, our system extracts features directly from raw data (instead of using pre-extracted features [Degottex et al., 2014a, Eyben et al., 2010a]). Our system consists of two modules (Fig 5.1): (1) the hierarchical encoder learns features at word-level and conversation-level;

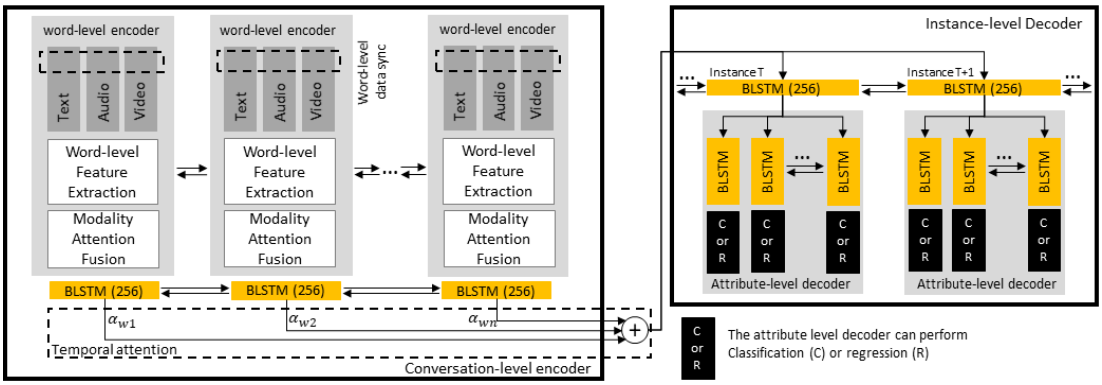


Figure 5.1: The structure of our proposed hierarchical encoder-decoder for conversation understanding.

(2) the hierarchical decoder makes continuous multi-label predictions at each time instance.

Hierarchical Encoder: (Fig 5.1, left). Our hierarchical encoder has two levels, the *word-level encoder (WLE)* and *conversation-level encoder (CLE)*. The *word-level encoder* (Fig 5.1, gray shaded region at left top) synchronizes and combines the features extracted from different sensors, and only selects informative information to form the shared representation. As different sensors have different sampling rates, we perform word-level data synchronization before feature extraction. Compared to multimodal frame-level [Zhang et al., 2017] and sentence-level encoding [Poria et al., 2017b], the *conversation-level encoder* (Fig 5.1, solid line to the left) combines useful information extracted about each word into a single feature vector. This encoder allows the system to make use of multimodal information extracted over the entire conversation to make the predictions. Because not all words are important, we use a temporal attention mechanism from neural machine translation to select the important word vectors [Bahdanau et al., 2014].

Hierarchical Decoder: (Fig 5.1, right). Our hierarchical decoder also has two levels, the *instance-level decoder (ILD)* and *attribute-level decoder (ALD)*. The *instance-level decoder* decodes the features to each time instance. During simple or short conversations, the emotion, sentiment, and traits remain the same, so we can decode all the information into a single instance. In such scenarios, the ILD treats the entire input data as a single time instance, performing a single prediction per case. During complex conversations with changing emotion, sentiment, and traits, the ILD allows us to make continuous predictions within a single conversation. In addition to multi-class classification, some datasets have more than one label per time instance (such as speaker trait analysis datasets). Our *attribute-level decoder* decodes the features at each time instance into multiple co-existing attributes and makes multi-label predictions. This hierarchical decoder structure can be applied to both classification and regression on both the instance-level and attribute-level.

5.3.2 Word-level Feature Extraction

Word-level feature synchronization and extraction are the foundations of our encoder. Word-level synchronization aligns the features by word, aiding fusion across different modalities. For datasets without word-level timestamps, we synchronized the audio and video to text using aeneas[2] from Sakoe-Chiba Band Dynamic Time Warping (DTW) [Sakoe and Chiba, 1978].

Even with word-level timestamps, sensors have different sampling rates that inhibit merging. We considered two options: (1) Downsample all data to a common rate, or (2) Extract features for each modality at their original sampling rate and encode them

[2]https://www.readbeyond.it/aeneas/

into a single vector. We adopted the second approach because downsampling loses information. We used different encoding strategies for different modalities (Fig 5.2):

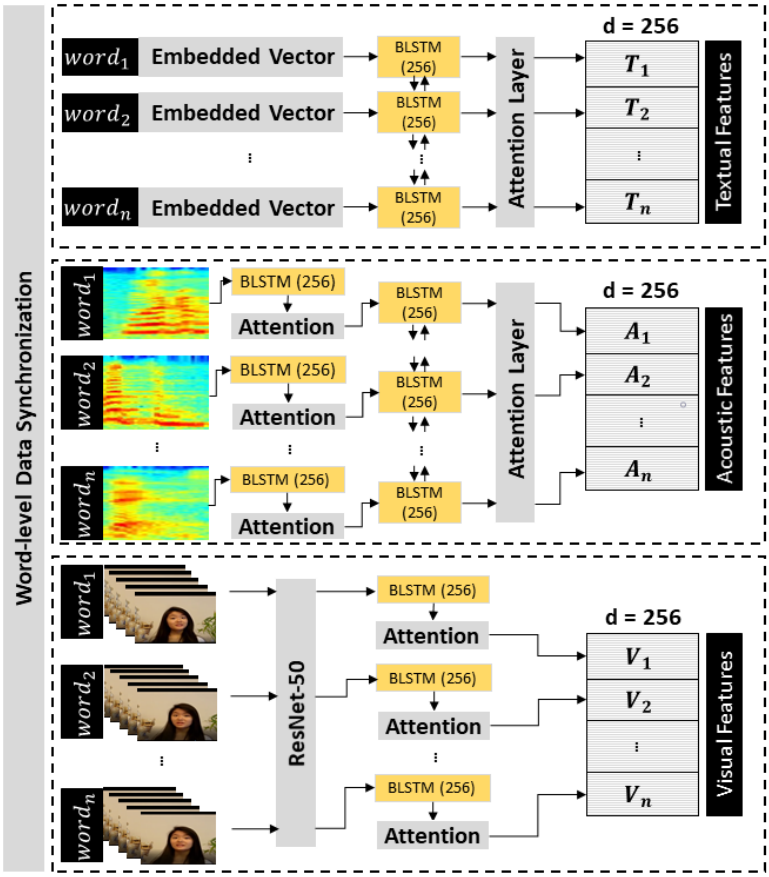


Figure 5.2: Word-level data synchronization and feature extraction using attention mechanisms.

Text: To capture textual word-level representations, we first embed each word using a pretrained *word2vec* dictionary [Mikolov et al., 2013]. We introduced an attention-based bidirectional LSTM (bi-LSTM) to extract word-level representations [Yang et al., 2016]. The key aspects of our approach are: (1) Instead of directly using the embedded vectors as textual features [Chen et al., 2017], our LSTM extracts high-level associations. (2) Instead of CNNs with fixed window size, our LSTM fully captures sequential information with varying lengths [Poria et al., 2015]. (3) The attention mechanism enables

selection of informative word-level representations [Zadeh et al., 2017]. We set the bi-LSTM dimension equal to 256 and the *word2vec* dimension as 300.

Audio: We introduce a hierarchical attention structure to extract informative acoustic features at both frame-level and word-level. Unlike previous research that directly used low-level acoustic descriptors (LLDs), we extracted 100 fps Mel-frequency spectral coefficients (MFSCs), which had been demonstrated effective on deep models due to locality maintenance and higher dimensionality [Gu et al., 2018a, Abdel-Hamid et al., 2014]. First, we use an attentive bi-LSTM (same as for the text feature extraction) to select the informative frames. The word-level representation is then a weighted sum of a word's frames. We then apply another attentive bi-LSTM over the word-level representations to learn the associations between representations and select informative representations at word-level. The final outputs are word-level acoustic representations with the same dimensionality as the word-level textual representations. As suggested in [Gu et al., 2018d], we used a 64-filter bank to extract MFSCs and initialized the bi-LSTM dimension as 256 at both frame-level and word-level.

Video: This branch captures the facial expression and body posture features from video. Previous work on facial recognition [Ranjan et al., 2017, Li et al., 2017] suggested that ResNet [He et al., 2016] performs well at person identification and tracking. We thus built our visual feature extractor with resnet-50 for each frame. Attention [Bahdanau et al., 2014] is applied to the extracted feature vectors for each frame to select the most representative features on each video frame for prediction. Because the datasets we used contains only one face at a time, we chose the Resnet for visual feature extraction. For more complex scenarios with multiple people per-frame, the face detector shall be implemented to capture features from certain face [Rosas et al., 2013].

Each input modality will generate feature vectors of the same length to avoid a dominant modality during fusion. The encoder structure finally outputs a visual, acoustic, and textual representation for each word.

5.3.3 Modality Attention and Fusion

To merge the word-level features for further processing, we introduce a novel modality fusion strategy. Modality fusion strategies are either early fusion or late fusion. Late fusion makes a final prediction from an ensemble of single-modality predictions (i.e. by voting); it is simple to implement but ignores associations between different modalities. Early fusion concatenates the features, but direct concatenation assumes that each modality contributes equally, which may not always be true (e.g. visual features are important for emotion recognition, but textual features contribute more to sentiment analysis). Attention is commonly used to select the most representative features.

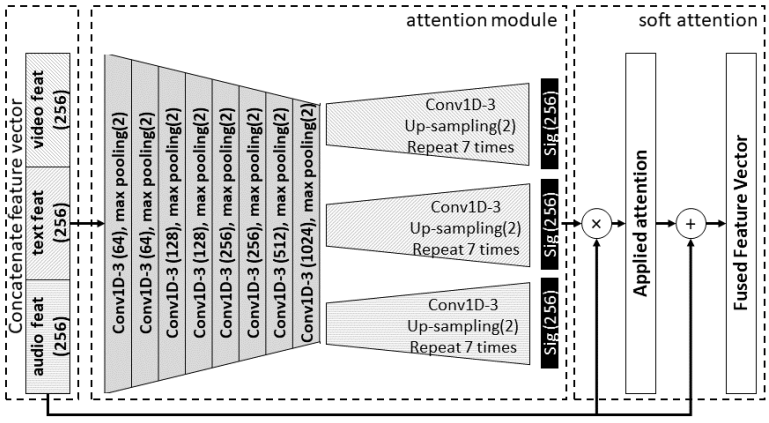


Figure 5.3: Our fusion strategy based on 1D fully convolutional network and soft attention.

Previous research focused on the attention over modality-specific features [Mirsamadi et al., 2017]

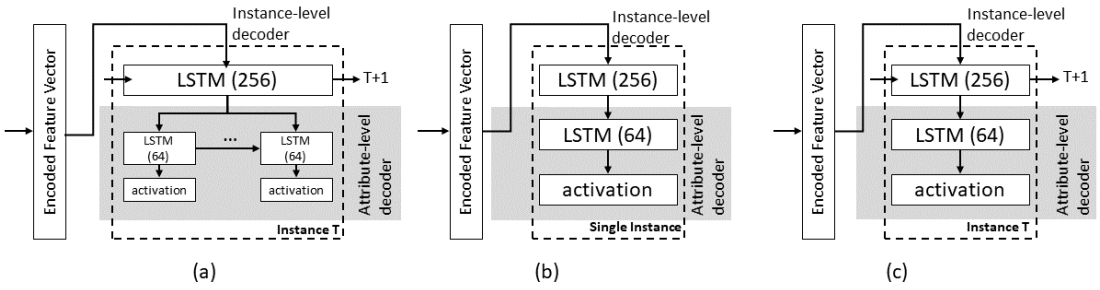
or attention over shared feature vectors [Chen et al., 2017, Poria et al., 2017c]. We propose a fusion structure with two-level attention: (1) a cross-modality attention focusing on the important modality; (2) a modality-specific attention that highlights important feature dimensions within each modality. We first select a modality by attention, and then expand it to modality-specific attention using a one-dimensional convolution-deconvolution network (Fig 5.3).

Our cross-modality attention is implemented with a set of convolution and pooling operations similar to visual attention used in image recognition [Wang et al., 2017]. Because convolution does not compromise spatial associations, the input vector for each modality will eventually be converted into a feature point with a high channel dimension (3×channel). This output denotes the importance of each input modality. This modality attention representation can be directly broadcasted to the input feature vector and used as the attention vector. However, such method overlooks the associations between modalities. We propose deconvolution and up-sampling (Fig 5.3) to create the attention vector of each input modality. Similarly, the attention vector of each modality is aligned to the modality's feature vector. Finally, we apply the attention to the feature vector input using soft attention [Wang et al., 2017].

5.3.4 Temporal Attention

We add attention for conversation-level feature fusion to select only the important word representations for final decision making. We adopted the temporal attention [Bahdanau et al., 2014] similar to the attentive bi-LSTM mechanism in textual feature extraction [Yang et al., 2016]. This enables the system to establish temporal association for both encoder and decoder between word-level representations (Fig 5.1, the region

labeled by dash line).



(a) The hierarchical decoder. (b) Modification 1: The hierarchical decoder with single time instance. (c) Modification 2: The hierarchical decoder with single attribute.

Figure 5.4: The decoder structure

5.3.5 The Decoder

The hierarchical decoder has two levels: instance-level and attribute-level. The instance-level decoder decodes the encoder output across time instances to make continuous predictions (Fig 5.4(a)). The proposed decoder can be further modified based on two specific requirements. By setting the number of time instances to one (Fig 5.4(b)) the decoder performs single per-case multi-label predictions. By changing the number of attributes to one (Fig 5.4(c)), the system makes multiple binary- or multi-class classification per-conversation.

The decoder can perform classification and regression regardless of the number of time instances or attributes. Classification is done by a single softmax activation in the output layer, subject to categorical cross-entropy loss. For regression, we scale all labels to (0, 1) and use one sigmoid neuron as the output, subject to mean-absolute error (MAE) loss.

5.4 Experiments

5.4.1 Implementation

We synchronized all modalities to text using aeneas. Due to our limited hardware resources, the feature extractors for text, audio, and video are pre-trained and we set all bi-LSTMs with 256 hidden states. We selected the ReLU activation function except for the attention layers. We initialized the learning rate as 0.01 and used the Adam optimizer. We also applied batch normalization and dropout function to address the overfitting issue. We used the 80-20 training and testing split across all modalities. The implementation is based on Keras with Tensorflow and is trained on two GTX1080Ti GPUs.

5.4.2 Dataset

We evaluated our model on five published datasets: two sentiment datasets (MOSI, MOUD), two emotion datasets (IEMOCAP, EmotiW), and one multi-label traits dataset (POM).

MOSI: A multimodal sentiment intensity and subjectivity dataset consisting of 93 English review videos with 2199 utterance segments [Zadeh et al., 2016]. We took the average score from five annotators as the ground-truth label (as in previous research [Zadeh et al., 2018]). Considering the speaker independence, there are 1755 training and 444 testing utterances. We used video, audio, and text for classification (binary-category and 7-category) and regression tasks (Fig 5.4(b, c)).

MOUD: A multimodal Spanish sentiment dataset including 79 videos with a *positive/negative* label for each of 498 utterances [Rosas et al., 2013]. Instead of translating

the sentences into English as previous research [Poria et al., 2017b], we randomly initialized the word vectors. Considering its small size, we only used it for generalization experiments. We used 59 videos for training and the remaining 20 for testing. Following previous research [Poria et al., 2017b], we removed the *neutral* label and kept only *positive/negative* labels.

IEMOCAP: The Interactive Emotional Dyadic Motion Capture database is an acted, multimodal, and multi-speaker dataset [Busso et al., 2008] containing ∼12 hours of video, speech, and text. For each sentence in the dataset, we took the voted results from different annotators as labels. We performed experiments on the 10-category configuration and tested the classification (Fig 5.4(a, b)) with this dataset.

EmotiW [3]: A multimodal audio-visual emotion recognition dataset. We used IBM Watson speech-to-text software[4] to transcribe the text data. We used the official training and evaluation set (we did not use the test set due to the lack of labels). We only used it for generalization experiments on the classification task (7-category) due to its small size.

POM: The Persuasion Opinion Multimodal dataset contains 904 movie review videos [Park et al., 2014]. Each video contains one speaker and is annotated with several traits. Following previous research [Zadeh et al., 2018], we used *confidence*, *passion*, *dominance*, *credibility*, *entertaining*, *reserved*, *trusting*, *relaxed*, *nervous*, *humorous*, and *persuasive* (11 multi-labels). To compare, we followed the same 700-204 training-testing split. We performed multiclass classification, regression, and multi-label classification (Fig 5.4(a, b, c)) on this dataset.

[3]https://cs.anu.edu.au/few/ChallengeDetails.html

[4]https://www.ibm.com/watson/developercloud/speech-to-text/api/v1/

5.5 Preliminary Results

5.5.1 Experimental Results and Comparison

We first compared our system with previous research using the same setup on three relatively large datasets [Zadeh et al., 2018]. The proposed method outperformed the previous state-of-the-art on all three datasets on different tasks (Table 5.1). Observe that: 1). Our system outperforms previous work with the same network configuration on different applications (emotion recognition, semantic analysis, and trait analysis). 2). The major differences between our system and previous state-of-the-art are the feature extraction and fusion strategies; our hierarchical encoder and modality attention help the system select representative features and result in higher performance (Table 5.2). Our proposed structure with attribute-level decoder achieves 37.9% accuracy on POM dataset which is the second-best result compared with the multiclass classification baselines (Table 5.1). The state-of-the-art system used 11 separate models for multiclass classification [Abdel-Hamid et al., 2014] while we used single model. Our model is more scalable and easier to be implemented compared with multi-binary model based solution.

5.5.2 Impact of Encoder Modalities and Attentions

We evaluated the importance of each modality during IEMOCAP and MOSI tests. We analyze how each of the following affects performance: having multimodal data, the proposed word-level encoder (WLE) with modality attention (MA), and the conversation-level encoder (CLE) with temporal attention (TA). By removing components of our

	MOSI(BC)		MOSI(MCC)	MOSI(MCR)	IEMOCAP(MCC)		POM(MCC)	POM(MLC)
	Acc.	F1-Score	Acc.	MAE	Acc.	F1-Score	Acc.	Acc.
Majority	50.2	0.501	17.5	1.864	21.2	0.074	24.0	\
RF[Breiman, 2001]	56.4	0.563	21.3	\	24.1	0.180	32.6	\
SVM[Cortes and Vapnik, 1995]	71.6	0.723	26.5	1.100	27.3	0.253	34.4	\
THMM[Morency et al., 2011]	50.7	0.454	17.8	\	23.5	0.108	23.8	\
C-MKL[Poria et al., 2015]	72.3	0.720	30.2	\	34.0	0.311	\	\
EF-HCRF[Quattoni et al., 2007]	65.3	0.654	24.6	\	32.0	0.205	\	\
MV-HCRF[Song et al., 2012]	65.6	0.657	24.6	\	32.0	0.205	\	\
DF[Nojavanasghari et al., 2016]	72.3	0.721	26.8	1.143	26.1	0.200	34.3	\
EF-LSTM[Zadeh et al., 2018]	73.3	0.732	32.4	1.023	34.1	0.323	36.4	\
MV-LSTM[Rajagopalan et al., 2016]	73.9	0.740	33.2	1.019	31.3	0.267	36.1	\
BC-LSTM[Poria et al., 2017b]	73.9	0.739	28.7	1.079	35.0	0.341	34.1	\
TFN[Zadeh et al., 2017]	74.6	0.745	28.7	1.040	36.0	0.345	31.9	\
MARN[Zadeh et al., 2018]	77.1	0.770	34.7	0.968	37.0	0.359	39.4	\
Ours	**77.5**	**0.774**	**38.5**	**0.932**	**39.4**	**0.383**	**39.6**	**37.9**

Table 5.1: Experimental results and comparison on MOSI, IEMOCAP, and POM. (BC) for binary classification, (MCC) for multiclass classification, (MCR) for multiclass regression, and (MLC) for multi-label classification (accuracy in percentage).

Method	IEMOCAP(MCC)		MOSI(MCC)	
	Acc.	F1	Acc.	F1
T	31.8	0.307	32.7	0.271
A	31.5	0.310	27.5	0.225
V	28.4	0.268	23.3	0.191
T+A	33.4	0.325	32.8	0.278
T+V	32.7	0.321	32.1	0.270
A+V	31.8	0.301	23.8	0.187
T+A+V (no MA & WLE)	34.9	0.341	35.2	0.301
T+A+V (no TA & CLE)	35.2	0.352	33.8	0.298
T+A+V (with all)	**39.4**	**0.383**	**38.5**	**0.331**

Table 5.2: Experimental results and comparison of modality importance (accuracy in percentage)

model, we were able to study the impact of different modalities and attentions (Table 5.2). We found that: 1). The multimodal structure with tri-modality achieves the highest accuracy and F1 score on IEMOCAP and MOSI, indicating that the different modalities indeed complement each other. 2). Text alone outperforms both video and audio on MOSI. However, text and audio have similar performance on IEMOCAP, indicating that vocal delivery is more important for emotion recognition than for sentiment analysis. 3). Removing modality attention (and directly using concatenated features for conversation-level encoding) causes a significant accuracy decrease on both datasets; -4.5% accuracy on IEMOCAP and -3.3% accuracy on MOSI (Table 5.2). This shows that the word-level encoder with modality attention has a positive influence. 4). Conversation-level encoding with temporal attention brings 4.2% and 4.7% accuracy improvement on IEMOCAP and MOSI (Table 5.2). This demonstrates that the temporal attention and hierarchical encoder also improves performance.

	MOSI(MCC)	IEMOCAP(MCC)	POM(MCC)
GRU	37.4	38.6	38.9
LSTM	37.9	39.1	39.3
bi-LSTM	38.5	39.4	39.6

Table 5.3: Encoder quantity analysis (accuracy in percentage)

	Acc. or MAE	Trainable Parameters	Training Time (s)
MCC	39.6	3.45×10^8	3.96×10^5
MLC	37.9	5.14×10^7	6.15×10^4
MCR	0.102	4.45×10^8	1.22×10^6
MLR	0.158	3.18×10^7	8.74×10^4

Table 5.4: Comparison of multiclass classification (MCC), multiclass regression (MCR), and multi-label classification (MLC), multi-label regression (MLR) on POM dataset (accuracy in percentage).

5.5.3 Impact of Recurrent Unit

We also made the quantity evaluation of different sequential models on the encoder structure. We did the baseline experiments that using GRUs and LSTMs as the encoder, respectively. The result in Table 5.3 indicates that the bi-LSTM as we used in the proposed architecture has the best performance on MOSI, IEMOCAP, and POM dataset.

5.5.4 Decoder Analysis

We evaluated the proposed decoder by comparing the different decision-making model performances on POM, IEMOCAP, and MOSI. Specifically, we tried to answer the following questions: 1). Is the proposed hierarchical decoder sufficiently flexible for different tasks? 2). How does multi-label classification and regression compare with baseline multiple-classifier solutions [Zadeh et al., 2018]? We found that: 1). The proposed architecture achieved state-of-the-art classification and regression on MOSI and POM (Table 5.1 and Table 5.4). The only difference between these two models is the

IEMOCAP (MCC) → EmotiW (MCC)		
Testing Set	Acc.	F1
IEMOCAP	39.4	0.383
EmotiW with[Poria et al., 2017b]	23.4	0.231
EmotiW with our method	26.1	0.244
MOSI (BC) → MOUD (BC)		
Testing Set	Acc.	F1
MOSI	77.5	0.774
MOUD with[Poria et al., 2017b]	52.7	\
MOUD with our method	59.5	0.592

Table 5.5: Experimental results on generalization (accuracy in percentage)

activation function; softmax for classification and hard sigmoid for regression. This shows that our decoder can handle both classification and regression very well. 2). Compared with MOSI and POM, IEMOCAP has emotional changes over the conversation. Nevertheless, our model achieved state-of-the-art (Table 5.1 and Table 5.2) on IEMOCAP simply by changing the number of decoding instances from 1 prediction per case to 100 predictions per case (one at each percentile). This shows our model can handle predictions over time, answering questions 1. 3). We further compared the performance of multiclass and multi-label [Liu and Chen, 2015] tasks on both regression and classification (Table 5.4). Using multiple classifiers (one for each attribute) still achieves 1.7% better performance but requires ∼7 times more parameters and ∼6.5 times more training time than our proposed single multi-label classifier. The proposed multi-label architecture still achieved performance comparable to many previous baselines (Table 5.1). To our best knowledge, we are the first to use multi-label classification and regression for conversation understanding, and our model shows the potential of multi-label decoders for conversation understanding tasks.

5.5.5 Generalization Test

Our system should be transferable enough to achieve good performance on one dataset despite being trained on another. We performed two sets of experiments to test the system's generalizability.

We first trained a model on the IEMOCAP emotion dataset and tested on both IEMOCAP and EmotiW (Table 5.5). To match the labels between the IEMOCAP and EmotiW, we changed *surprise* to *exciting* due to the lack of *surprise* in IEMOCAP. Then, we trained our model on the MOSI sentiment dataset and tested on MOUD (Table 5.5). Similar to previous experiments [Zadeh et al., 2018], we processed the MOUD labels to positive/negative to perform transfer learning from the MOSI-trained model.

The results show that training on IEMOCAP achieves 26.1% accuracy with 0.244 F1 score on EmotiW. We further compared the generalization ability of our method with previous approaches [Poria et al., 2016]. We re-implemented their method with the same training and testing split. The proposed system outperforms the previous approach by 2.7% accuracy. The MOSI-trained model achieves 59.5% accuracy with 0.592 F1 score on MOUD. Compared to the 77.5% accuracy with 0.774 F1 score on the MOSI test set, the lower generalization accuracy might be caused by the language difference (MOSI is English, but MOUD is Spanish). However, compared with the approach in [Poria et al., 2017b], our system still outperforms the previous method by 6.8% accuracy.

5.5.6 Visualization of Attentions

Unlike previous approaches using uninterpretable attention at the hidden-state-level [Zadeh et al., 2018] or feature-level [Poria et al., 2017c], our system provides direct visualizations of modality and temporal attention. The word-level design allows intuitive understanding of modality importance of each word and its importance in the text sequence.

We noticed that the model assigns the text modality a high attention score on the words that carry emotion information, such as "dumb" (Figure 5). The system also focuses on audio or video when tone or facial expression changes. For example, the system focuses on video when the kid has knitted brows and then on audio when he has a higher acoustic energy distribution (Fig 5.5(a)). The visualization demonstrates that the modality attention selects useful information from different sources.

Unlike modality attention, temporal attention assigns the importance of each word representation. We see that the temporal attention vector peaks on "great", indicating this representation contains the most informative information (Fig 5.5(b)). Because temporal attention is computed over a multimodal representation, it selects the most important words considering all modalities.

5.6 Future Work

Our hierarchical decoder can be used for text and conversation understanding. It is common that a conversation or text has multiple attributes: a person can be both *angry* and *sad*, *excited* and *happy*. Unlike previous strategies [Zadeh et al., 2018], our hierarchical encoder-decoder framework flexibly decodes one or multiple attributes simultaneously over time with a single model. In addition, our temporal and modality

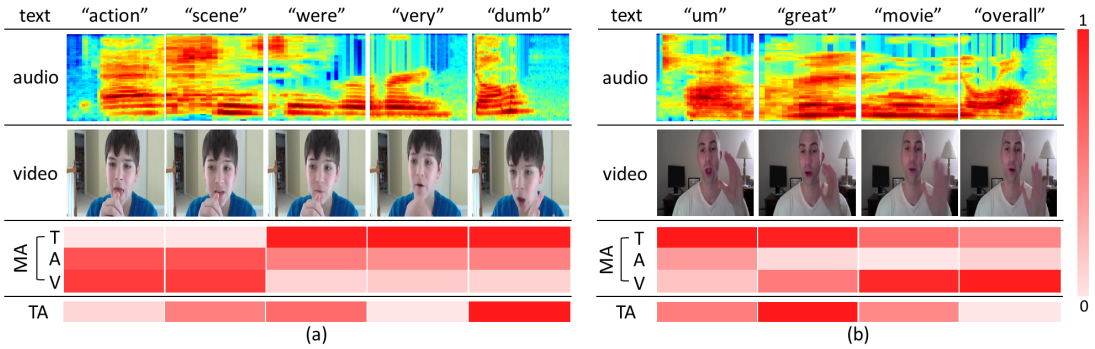


Figure 5.5: Visualization of modality attention (MA) and temporal attention (TA) on MOSI. (a) Negative example. (b) Positive example.

attentions allow visual interpretation of the model's inner workings (Fig 5.5).

We demonstrated that our hierarchical decoder is able to make both multiclass and multi-label predictions with performance comparable to most of previous systems (Table 5.1). Yet there are still some limitations: 1). the attribute-decoder framework assumes associations are between conversational attributes: e.g. if a conversation is convincing, it is usually pervasive. However, such complicated associations cannot be fully captured through limited data size (only 904 samples) with an imbalanced amount of data across attributes (e.g. there is a lack of labels with *confidence* score). We will continue to test our system on different datasets with concurrent labels. 2). The model simply decodes the entire conversation into time instances with equal length, which cannot make sentence-level predictions as in previous research. This may lead to some performance mismatches between our system and the previous approaches. We plan to decode features for sentence-level instance predictions in the future.

5.7 Summary

We introduced a novel human conversation analysis system using a hierarchical encoder-decoder framework. To better combine features extracted from different modalities, we proposed word-level fusion with modality attention. Our system achieved state-of-the-art performance on three published datasets and outperformed others at generalization testing. We hope to deliver the following contributions to the community:

1. A hierarchical encoder-decoder framework that can recognize emotion, sentiment, and speaker traits.

2. A word-level feature extraction strategy that can be widely used for emotion recognition, sentiment analysis, and associated applications.

3. An attentive modality fusion strategy that can be used for any multimodal application.

4. A detailed comparison with previous work for future reference, including generalization and classification vs. regression tests.

5. Datasets synchronized to the word-level and our source code for future comparative research.

Chapter 6

Conclusion

The speech affective computing is one of the most popular research topic in artificial intelligent. In this research, we mainly focus on improving the model and system performance of speech emotion recognition and sentiment analysis on three aspects: modality-specific feature extraction, modality fusion, and context-aware design.

For feature extraction, we first evaluate the performance low-level handcraft features and high-level features extracted by the deep neural network. Then, we propose the hierarchical attention network to extract both the acoustic and textual features on word-level. We further improve the model by introducing the sub-view attention module. Our contribution on affective feature extraction can be summarized as:

1. A hybrid architecture using attention mechanism with recurrent neural network to select the informative acoustic features and textual features independently.

2. A hierarchical attention structure to represent the acoustic and textual features on word-level.

3. A sub-view attention module that only relies on the attention mechanism to extract modality-specific features without using recurrent neural networks.

For modality fusion, we first present a hybrid structure that combines low-level features with high-level features using a deep neural network. Then, we improve the

fusion by applying the modality attention to help the model to select the helpful features on modality-level. We further introduce three attention based fusion strategies to combine modality-specific features on word-level. We also design a mutual correlation attentive factor to help the sub-view attention architecture to learn the across modality association. Our contribution can be summarized as:

1. A modality attention module that generates scores to represent the importance for each modality and fuses the features with the weighted scores.

2. Three word-level fusion strategies to combine features and learn correlations in a common time scale across different modalities.

3. A novel mutual correlation attentive factor that automatically learns the associations across modalities in each sub-view attention layer to facilitate fusion.

For context-aware design, we introduce the mutual attentive fusion network that uses the attention to give weighted scores to all the previous utterances as the context representation. We further apply the encoder-decoder structure to generate the affective state based on continues input. We also propose a solution for the disagreeing annotation, which is an important issue of most existing context-aware systems. The contribution can be summarized as:

1. An efficient dyadic fusion network that mainly relies on an attention mechanism for feature extraction, modality fusion, and contextual representation learning.

2. A hierarchical encoder-decoder framework. The encoder encodes features from low-level to high-level. The decoder first decodes the abstract features into attribute profile at each time instance, and then decodes the attribute profile at

each time instance to individual attributes.

3. An effective solution and a detailed experimental analysis of the label disagreement issue that keeps sequence consistency and allows full use of labeled dialog data.

References

[Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.

[Abdel-Hamid et al., 2014] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545.

[Badjatiya et al., 2017] Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Balazs and Velásquez, 2016] Balazs, J. A. and Velásquez, J. D. (2016). Opinion mining and information fusion: a survey. *Information Fusion*, 27:95–110.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

[Busso et al., 2013] Busso, C., Bulut, M., Narayanan, S., Gratch, J., and Marsella, S. (2013). Toward effective automatic recognition systems of emotion in speech. *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pages 110–127.

[Cai and Xia, 2015] Cai, G. and Xia, B. (2015). Convolutional neural networks for multimedia sentiment analysis. In *Natural Language Processing and Chinese Computing*, pages 159–167. Springer.

[Cambria, 2016] Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

[Chen et al., 2017] Chen, M., Wang, S., Liang, P. P., Baltrušaitis, T., Zadeh, A., and Morency, L.-P. (2017). Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171. ACM.

[Chen et al., 2018] Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Ku, L.-W., et al. (2018). Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

[Chollet et al., 2015] Chollet, F. et al. (2015). Keras.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Degottex et al., 2014a] Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. (2014a). Covarep-a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE.

[Degottex et al., 2014b] Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. (2014b). Covarep—a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE.

[El Ayadi et al., 2011] El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

[Eyben et al., 2010a] Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., and Cowie, R. (2010a). On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3(1-2):7–19.

[Eyben et al., 2010b] Eyben, F., Wöllmer, M., and Schuller, B. (2010b). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

[Forbes-Riley and Litman, 2004] Forbes-Riley, K. and Litman, D. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

[Giles and Powesland, 1975] Giles, H. and Powesland, P. F. (1975). *Speech style and social evaluation*. Academic Press.

[Gu et al., 2018a] Gu, Y., Chen, S., and Marsic, I. (2018a). Deep multimodal learning for emotion recognition in spoken language. *arXiv preprint arXiv:1802.08332*.

[Gu et al., 2017a] Gu, Y., Li, X., Chen, S., Li, H., Farneth, R. A., Marsic, I., and Burd, R. S. (2017a). Language-based process phase detection in the trauma resuscitation. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 239–247. IEEE.

[Gu et al., 2017b] Gu, Y., Li, X., Chen, S., Zhang, J., and Marsic, I. (2017b). Speech intention classification with multimodal deep learning. In *Canadian Conference on Artificial Intelligence*, pages 260–271. Springer.

[Gu et al., 2018b] Gu, Y., Li, X., Huang, K., Fu, S., Yang, K., Chen, S., Zhou, M., and Marsic, I. (2018b). Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 537–545. ACM.

[Gu et al., 2019] Gu, Y., Lyu, X., Sun, W., Li, W., Chen, S., Li, X., and Marsic, I. (2019). Mutual correlation attentive factors in dyadic fusion networks for speech emotion recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 157–166.

[Gu et al., 2018c] Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., and Marsic, I. (2018c). Hybrid attention based multimodal network for spoken language classification. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 2379. NIH Public Access.

[Gu et al., 2018d] Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., and Marsic, I. (2018d). Multimodal affective analysis using hierarchical attention strategy with word-level alignment. *arXiv preprint arXiv:1805.08660*.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

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

[Hu et al., 2013] Hu, X., Tang, J., Gao, H., and Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618. ACM.

[Huang and Narayanan, 2016] Huang, C.-W. and Narayanan, S. S. (2016). Attention assisted discovery of sub-utterance structure in speech emotion recognition. In *INTERSPEECH*, pages 1387–1391.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456.

[Jin et al., 2015] Jin, Q., Li, C., Chen, S., and Wu, H. (2015). Speech emotion recognition with acoustic and lexical features. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4749–4753. IEEE.

[Kim et al., 2007] Kim, S., Georgiou, P. G., Lee, S., and Narayanan, S. (2007). Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pages 48–51. IEEE.

[Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[Kim et al., 2000] Kim, Y.-H., Hahn, S.-Y., and Zhang, B.-T. (2000). Text filtering by boosting naive bayes classifiers. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 168–175. ACM.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Kishore and Satish, 2013] Kishore, K. K. and Satish, P. K. (2013). Emotion recognition in speech using mfcc and wavelet features. In *2013 3rd IEEE International Advance Computing Conference (IACC)*, pages 842–847. IEEE.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[Lee and Tashev, 2015] Lee, J. and Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[Li et al., 2017] Li, X., Zhang, Y., Zhang, J., Chen, Y., Li, H., Marsic, I., and Burd, R. S. (2017). Region-based activity recognition using conditional gan. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1059–1067. ACM.

[Liang et al., 2018a] Liang, P. P., Liu, Z., Zadeh, A., and Morency, L.-P. (2018a). Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920*.

[Liang et al., 2018b] Liang, P. P., Zadeh, A., and Morency, L.-P. (2018b). Multimodal local-global ranking fusion for emotion recognition. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 472–476. ACM.

[Liscombe et al., 2005] Liscombe, J., Riccardi, G., and Hakkani-Tur, D. (2005). Using context to improve emotion detection in spoken dialog systems.

[Litman and Forbes-Riley, 2004] Litman, D. J. and Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 351. Association for Computational Linguistics.

[Liu and Chen, 2015] Liu, S. M. and Chen, J.-H. (2015). A multi-label classification based approach for sentiment classification. *Expert Systems with Applications*, 42(3):1083–1093.

[Liu et al., 2018] Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., and Morency, L.-P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

[Luengo et al., 2005] Luengo, I., Navas, E., Hernáez, I., and Sánchez, J. (2005). Automatic emotion recognition using prosodic parameters. In *Ninth European Conference on Speech Communication and Technology*.

[Melville et al., 2009] Melville, P., Gryc, W., and Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284. ACM.

[Metallinou et al., 2012] Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., and Narayanan, S. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Mirsamadi et al., 2017] Mirsamadi, S., Barsoum, E., and Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2227–2231. IEEE.

[Mishne et al., 2005] Mishne, G. et al. (2005). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, volume 19, pages 321–327.

[Morency et al., 2011] Morency, L.-P., Mihalcea, R., and Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176. ACM.

[Murray and Arnott, 1993] Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108.

[Neumann and Vu, 2017] Neumann, M. and Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv preprint arXiv:1706.00612*.

[Nojavanasghari et al., 2016] Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., and Morency, L.-P. (2016). Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288. ACM.

[Nwe et al., 2003] Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623.

[Oneto et al., 2016] Oneto, L., Bisio, F., Cambria, E., and Anguita, D. (2016). Statistical learning theory and elm for big social data analysis. *ieee CompUTATionAl inTelliGenCe mAGAzine*, 11(3):45–55.

[Park et al., 2014] Park, S., Shim, H. S., Chatterjee, M., Sagae, K., and Morency, L.-P. (2014). Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57. ACM.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Poria et al., 2017a] Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017a). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

[Poria et al., 2015] Poria, S., Cambria, E., and Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544.

[Poria et al., 2017b] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. (2017b). 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)*, volume 1, pages 873–883.

[Poria et al., 2017c] Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A., and Morency, L.-P. (2017c). Multi-level multiple attentions for contextual multimodal sentiment analysis. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 1033–1038. IEEE.

[Poria et al., 2016] Poria, S., Chaturvedi, I., Cambria, E., and Hussain, A. (2016). Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 439–448. IEEE.

[Poria et al., 2018] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

[Quattoni et al., 2007] Quattoni, A., Wang, S., Morency, L.-P., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 29(10).

[Rajagopalan et al., 2016] Rajagopalan, S. S., Morency, L.-P., Baltrusaitis, T., and Goecke, R. (2016). Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*, pages 338–353. Springer.

[Ranjan et al., 2017] Ranjan, R., Patel, V. M., and Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Ringeval et al., 2015] Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D., and Schuller, B. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30.

[Rosas et al., 2013] Rosas, V. P., Mihalcea, R., and Morency, L.-P. (2013). Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45.

[Rozgic et al., 2012] Rozgic, V., Ananthakrishnan, S., Saleem, S., Kumar, R., and Prasad, R. (2012). Ensemble of svm trees for multimodal emotion recognition. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4. IEEE.

[Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

[Savran et al., 2012] Savran, A., Cao, H., Shah, M., Nenkova, A., and Verma, R. (2012). Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 485–492. ACM.

[Schuller, 2011] Schuller, B. (2011). Recognizing affect from linguistic information in 3d continuous space. *IEEE Transactions on Affective computing*, 2(4):192–205.

[Schuller et al., 2005] Schuller, B., Müller, R., Lang, M., and Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Ninth European Conference on Speech Communication and Technology*.

[Schuller et al., 2004] Schuller, B., Rigoll, G., and Lang, M. (2004). 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*, volume 1, pages I–577. IEEE.

[Seppi et al., 2008] Seppi, D., Batliner, A., Schuller, B., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., and Aharonson, V. (2008). Patterns, prototypes, performance: classifying emotional user states. In *Ninth Annual Conference of the International Speech Communication Association*.

[Socher et al., 2013] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

[Song et al., 2012] Song, Y., Morency, L.-P., and Davis, R. (2012). Multi-view latent variable discriminative models for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2120–2127. IEEE.

[Sun et al., 2017] Sun, S., Luo, C., and Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25.

[Tokozume et al., 2017] Tokozume, Y., Ushiku, Y., and Harada, T. (2017). Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv:1711.10282*.

[Toutanova et al., 2003] Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pages 173–180. Association for Computational Linguistics.

[Trigeorgis et al., 2016] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE.

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

[Wang et al., 2017] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. (2017). Residual attention network for image classification. *arXiv preprint arXiv:1704.06904*.

[Wang et al., 2014] Wang, F., Wang, Z., Li, Z., and Wen, J.-R. (2014). Concept-based short text classification and ranking. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1069–1078. ACM.

[Wang et al., 2016] Wang, H., Meghawat, A., Morency, L.-P., and Xing, E. P. (2016). Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*.

[Wang and Tashev, 2017] Wang, Z.-Q. and Tashev, I. (2017). Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5150–5154. IEEE.

[Wöllmer et al., 2013a] Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., and Rigoll, G. (2013a). Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163.

[Wöllmer et al., 2013b] Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L.-P. (2013b). Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

[Wu and Tsai, 2014] Wu, C.-E. and Tsai, R. T.-H. (2014). Using relation selection to improve value propagation in a conceptnet-based sentiment dictionary. *Knowledge-Based Systems*, 69:100–107.

[Wu and Liang, 2010] Wu, C.-H. and Liang, W.-B. (2010). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21.

[Yang et al., 2016] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

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

[Zadeh et al., 2018] Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., and Morency, L.-P. (2018). Multi-attention recurrent network for human communication comprehension. *arXiv preprint arXiv:1802.00923*.

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

[Zhang et al., 2017] Zhang, S., Zhang, S., Huang, T., Gao, W., and Tian, Q. (2017). Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

[Zheng et al., 2015] Zheng, W., Yu, J., and Zou, Y. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 827–831. IEEE.