Mutual Correlation Attentive Factors in Dyadic Fusion Networks for Speech Emotion Recognition

Yue Gu
Rutgers University
yg202@rutgers.edu

Xinyu Lyu
Rutgers University
xl422@rutgers.edu

Weijia Sun
Rutgers University
freddie.sun@rutgers.edu

Weitian Li
Rutgers University
wl436@rutgers.edu

Shuhong Chen
Rutgers University
sc1624@rutgers.edu

Xinyu Li
Amazon Inc.
Rutgers University
xl264@rutgers.edu

Marsic Ivan
Rutgers University
marsic@rutgers.edu

ABSTRACT

Emotion recognition in dyadic communication is challenging because: 1. Extracting informative modality-specific representations requires disparate feature extractor designs due to the heterogenous input data formats. 2. How to effectively and efficiently fuse unimodal features and learn associations between dyadic utterances are critical to the model generalization in actual scenario. 3. Disagreeing annotations prevent previous approaches from precisely predicting emotions in context. To address the above issues, we 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. Our approach has three distinct characteristics: 1. Instead of using a recurrent neural network to extract temporal associations as in most 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: IEMOCAP and MELD. Our model significantly outperforms previous state-of-the-art research by 3.8%-7.5% accuracy, using a more efficient model.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351039

CCS CONCEPTS

• Computing methodologies → Artificial intelligence → Natural language processing

• Information systems → Information systems applications → Multimedia information systems → Multimedia streaming

KEYWORDS

Speech Emotion Recognition, Dyadic Communication, Multimodal Fusion Network, Attention Mechanism, Mutual Correlation Attentive Factor.

ACM Reference format:

Yue Gu, Xinyu Lyu, Weijia Sun, Weitian Li, Shuhong Chen, Shuhong Xinyu Li, and Ivan Marsic. 2019. Mutual Correlation Attentive Factors in Dyadic Fusion Networks for Speech Emotion Recognition. In *2019 ACM Multimedia Conference (MM '19), October 21–25, 2019, Nice, France.* 9 pages. https://doi.org/10.1145/3343031.3351039.

1. INTRODUCTION

Speech emotion recognition, which aims to automatically identify emotional state from human verbal expression, has become a rapidly expanding research topic with the development of social media technology. Precisely detecting human emotion is useful in many real-world applications such as recommender systems and chatbots. 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 [1-3]. 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 [4-6]. 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.

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 [2, 7]. Using the appropriate approaches 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 [6, 8]. 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 [9], 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 [1-7]. 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 [10] 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 [6, 11], 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 [4, 5, 11]. 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 tested our model on two published multimodal emotion recognition datasets: IEMOCAP [9] and MELD [12]. 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. The main contribution of our paper 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 novel mutual correlation attentive factor that automatically learns the associations across modalities in each sub-view attention layer to facilitate fusion.
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.

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 handcrafted features and abstract high-level representations. A large body of low-level features for both the text and audio branches 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 [13-16]. 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 [2, 4, 17]. The long short-term memory network was applied to both the text and audio branch to capture the temporal features [4, 18, 31]. More recently, attention mechanisms were integrated with recurrent neural networks to select informative textual and acoustic features [5, 6, 11]. 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 [3, 8] to further improve modality-specific feature extraction. Most previous works focused on using single utterance to identity emotion [2, 3], while the more recent works started combining the surrounding utterances as context to provide extra information for utterance-level emotion recognition [4-6]. 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 [19], 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 [1, 7, 20], which allows significant performance improvement. To further measure the temporal and context information, a multi-attention recurrent network was proposed [5] 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 [21, 32] 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 [6]. 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. The rest of the paper is organized as follows: We provide the model architecture details in section 3. Section 4 presents the results, which we discuss in section 5. Finally, section 6 concludes the paper.

3. METHODOLOGY

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.

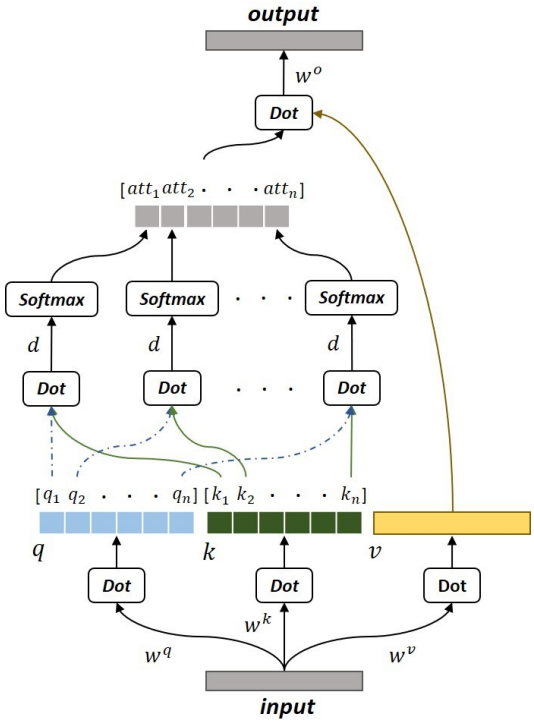

Figure 1: Sub-view attention mechanism

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 [10] 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 [10]. 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. 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 ($w^q, w^k,$ and w^v), respectively:

$$q, k, v = linear(iw^q, iw^k, iw^v) \quad (1)$$

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:

$$att_j = softmax\left(\frac{q_j k_j^{\mathrm{T}}}{\sqrt{d}}\right), \ j \in [1, n] \quad (2)$$

where d is the scale dimension and $n \cdot d$ 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:

$$o = [concate(att_1, att_2, \dots, att_n)v]w^o \quad (3)$$

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.

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 [4, 6] 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. 2, we first embed each word into a 200-dimensional vector using pretrained word vectors from *Glove* [22]. Then, we feed the embedded word vectors into 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 [15] 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. 2, the acoustic representation for each utterance is also a 200-dimensional vector.

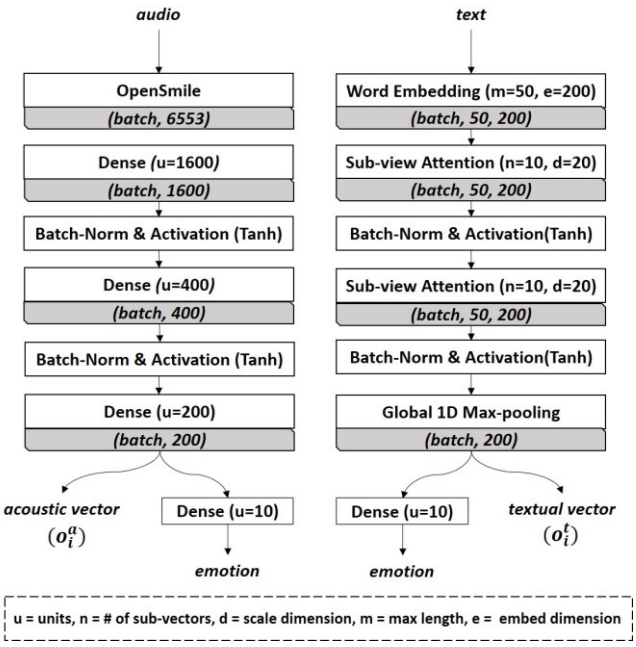

Figure 2: Modality-specific feature extraction

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.

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 [4, 5], 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.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^tor 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:

$$k^{t^*} = k^t + l^t k^a \quad (4)$$

$$k^{a^*} = k^a + l^t k^t \quad (5)$$

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. 3 shows the details of the mutual correlation attentive factor.

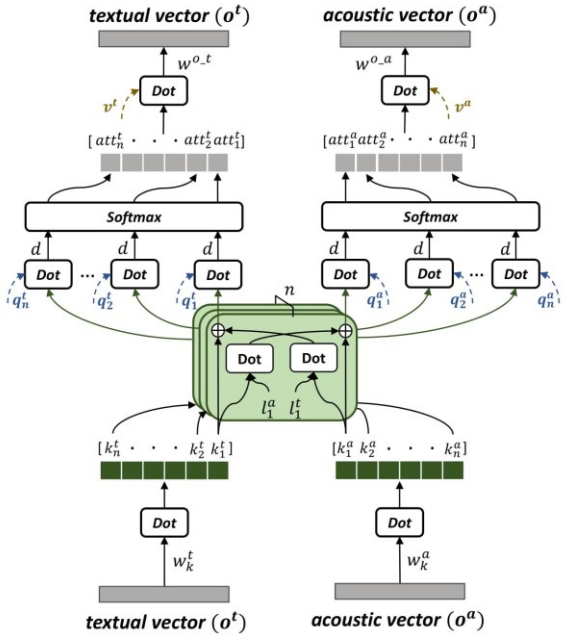

Figure 3: Mutual correlation attentive factors (MCAF) in sub-view attention for modality fusion

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.

3.5 Decision Making

Fig. 4 shows the overall structure of the dyadic fusion network. As suggested by the self-attention mechanism [10], we first connect each MCAF sub-view attention layer with a batch normalization layer [23] 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).

Compared to the previous approaches that classify emotion only based on all-agreeing or majority-voted labels [1-4], 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.

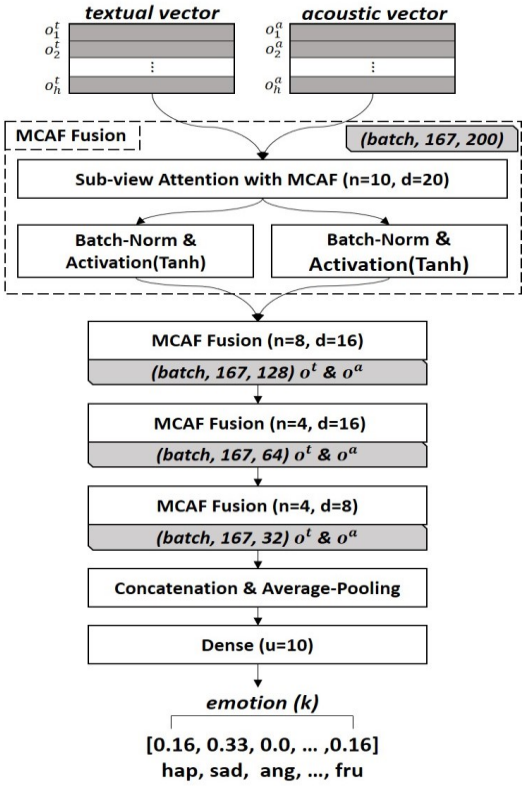

Figure 4: Dyadic fusion network

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 [6], 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 [24]. We finally compute the argmax based on the output to transform the regression metric into a categorical metric.

4. EXPERIMENT

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 [9]. 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 [12, 25]. 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.

4.2 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 [26].

RF: a random forest model that also uses the concatenated text and audio branch features [27].

C-MKL: a convolutional neural network with a multiple kernel learning strategy to predict emotion and sentiment based on multimodal data [7].

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 [5].

BC-LSTM: a context-dependent model using two unidirectional LSTMs to predict human sentiment and emotion, which can identify information from context utterances [4].

MV-LSTM: a recurrent model to capture both modality-specific and cross-view interactions over time or structured outputs from multiple modalities [18].

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 [1].

HAW: a multimodal structure using hierarchical attention with word-level alignment to utterance-level sentiment and emotion [3].

AMN: an attentive multimodal network using a hierarchical encoder-decoder to predict the sentiment and emotions with contextual information [6].

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) [5].

4.3 Implementation

We implement the model with *Keras* [28] and *Tensorflow* [29] 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. 2 and Fig. 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 [30]. 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 [5]. 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 [6], We assign the disagreeing labels to the '*other*' category for all baselines.

Table 1: Emotion recognition result on IEMOCAP dataset. Following previous research, the metric computation based on 9 categories (without '*other*').

	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**

5. RESULT AND DISCUSSION

5.1 Comparison with Baselines

We first compare our model with the baselines and the state-of-the-art on IEMOCAP. Following previous research [5], 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 1: 1. The significant performance improvement shows that the proposed architecture is effective for multi-class classification.

Table 2: Emotion recognition result on MELD dataset (%). the metric computation based on binary classification for each emotion. Ang=anger, Neu=neutral, Sur=surprise.

	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(A)	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**

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.

Table 3: Quantitative analysis on IEMOCAP dataset (%). Ang=anger, Neu=neutral+frustration, Hap=happy+exciting.

	Acc	Weighed-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**	**50.3**	31.8	**54.1**	**74.1**	**61.0**

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 [12]. Due to the imbalanced emotion split in MELD, we conduct binary classification during experiments. The result in Table 3 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.

5.2 Quantitative Analysis

We further evaluate our model by comparing the performance of unimodal and multimodal structures (shown in Table 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 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 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 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 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.

Table 4: Comparison of training cost on IEMOCAP dataset.

	Trainable Parameters	Training FLOPs	Training Speed (ms/per epoch)	Acc. (%)
BC-LSTM	$2.9 \cdot 10^7$	$5.7 \cdot 10^7$	$2.3 \cdot 10^4$	38.9
Ours	$1.3 \cdot 10^7$	$2.7 \cdot 10^7$	$1.2 \cdot 10^4$	51.6

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 [1-8]. In this section, we provide an analysis of the samples that cannot be assigned to a category in IEMOCAP.

Table 5: Analysis of disagreement annotation on IEMOCAP dataset.

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

# of disagreement annotation utterance/total utterances

	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

P/A = number of the positive samples / number of total samples. AP = average precision.

As shown in Table 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

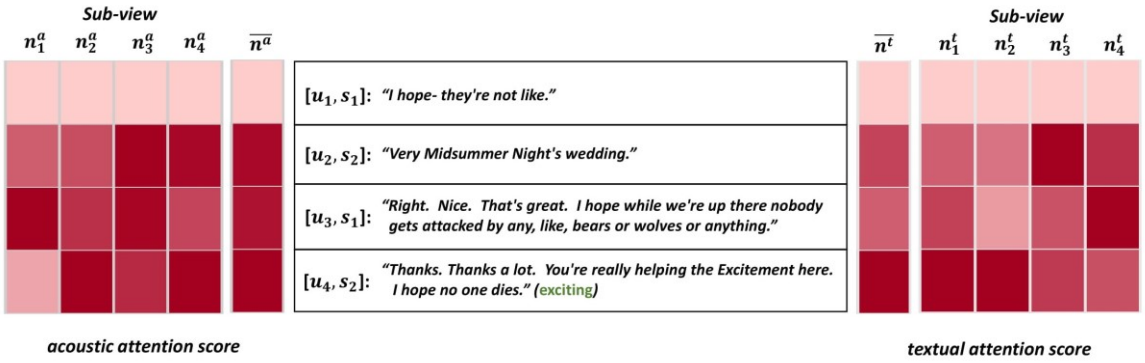

Figure 5: Attention visualization. $\overline{n^t}$, $\overline{n^a}$: average sub-view scores. n_i^t, n_i^a : sub-view attention score. u_i : the index of the utterances. s_1, s_2 : speaker IDs.

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 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.

5.4 Attention Visualization

In this section, we provide an example of the sub-view attention in Fig.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 represent the importance of the current utterance. As shown in Fig. 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.

6. CONCLUSIONS

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.

ACKNOWLEDGMENTS

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported in part by the National Institutes of Health under Award Number R01LM011834.

REFERENCES

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

[2] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 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*, pp. (2015). 2539-2544.

[3] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. "Multimodal affective analysis using hierarchical attention strategy with word-level alignment." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 2225--2235).

[4] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. "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)*, pp. 873-883.

[5] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. "Multi-attention recurrent network for human communication comprehension." In *Thirty-Second AAAI Conference on Artificial Intelligence*. (2018).

[6] Yue Gu, Xinyu Li, Kaixiang Huang, Shiyu Fu, Kangning Yang, Shuhong Chen, Moliang Zhou, and Ivan Marsic. 2018. "Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder." In *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 537-545. ACM, (2018).

[7] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. "Convolutional MKL based multimodal emotion recognition and sentiment analysis." In *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 439-448. IEEE, (2016).

[8] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. "Multimodal sentiment analysis with word-level fusion and reinforcement learning." In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 163-171. ACM, (2017).

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. "IEMOCAP: Interactive emotional dyadic motion capture database." *Language resources and evaluation* 42, no. 4 (2008): 335.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008. (2017).

[11] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. "Multi-level multiple attentions for contextual multimodal sentiment analysis." In *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 1033-1038. IEEE, (2017).

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

[13] Dino Seppi, Anton Batliner, Björn Schuller, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, and Vered Aharonson. 2008. "Patterns, prototypes, performance: classifying emotional user states." In Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia, pp. 601-604. (2008).

[14] Arman Savran, Houwei Cao, Miraj Shah, Ani Nenkova, and Ragini Verma. 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*, pp. 485-492. ACM, (2012).

[15] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. "Opensmile: the munich versatile and fast open-source audio feature extractor." In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459-1462. ACM, (2010).

[16] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. "COVAREP—A collaborative voice analysis repository for speech technologies." In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 960-964. IEEE, (2014).

[17] Yue Gu, Shuhong Chen, and Ivan Marsic. 2018. "Deep Mul timodal learning for emotion recognition in spoken language." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5079-5083. IEEE, (2018).

[18] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. "Extending long short-term memory for multi-view structured learning." In *European Conference on Computer Vision*, pp. 338-353. Springer, Cham, (2016).

[19] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. "Youtube movie reviews: Sentiment analysis in an audio-visual context." *IEEE Intelligent Systems* 28, no. 3 (2013): 46-53.

[20] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. "Efficient low-rank multimodal fusion with modality-specific factors." *arXiv preprint arXiv:1806.00064* (2018).

[21] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. "Multimodal local-global ranking fusion for emotion recognition." In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pp. 472-476. ACM, (2018).

[22] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543. (2014).

[23] Sergey Ioffe, and Christian Szegedy. 2015. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167* (2015).

[24] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. 2017. "Learning from between-class examples for deep sound recognition." *arXiv preprint arXiv:1711.10282* (2017).

[25] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, and Lun-Wei Ku. 2018. "Emotionlines: An emotion corpus of multi-party conversations." *arXiv preprint arXiv:1802.08379* (2018).

[26] Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. "Multimodal sentiment analysis of spanish online videos." *IEEE Intelligent Systems* 28, no. 3 (2013): 38-45.

[27] Leo Breiman. 2001. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32.

[28] François Chollet. 2015. "Keras." (2015).

[29] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin et al. 2016. "Tensorflow: A system for large-scale machine learning." In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265-283. (2016).

[30] Diederik P. Kingma, and Jimmy Ba. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[31] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. "Hybrid Attention based Multimodal Network for Spoken Language Classification." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2379-2390. (2018).

[32] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. "Multimodal language analysis with recurrent multistage fusion." arXiv preprint arXiv:1808.03920 (2018).