Progressive Modality Reinforcement for Human Multimodal Emotion Recognition from Unaligned Multimodal Sequences

Fengmao Lv[1,2] Xiang Chen[3] Yanyong Huang[2*] Lixin Duan[4] Guosheng Lin[5*]

[1] Southwest Jiaotong University
[2] Center of Statistical Research, Southwestern University of Finance and Economics
[3] Platform and Content Group, Tencent
[4] University of Electronic Science and Technology of China
[5] Nanyang Technological University

fengmaolv@126.com {chx245296678,lxduan}@gmail.com huangyy@swufe.edu.cn gslin@ntu.edu.sg

Abstract

Human multimodal emotion recognition involves time-series data of different modalities, such as natural language, visual motions, and acoustic behaviors. Due to the variable sampling rates for sequences from different modalities, the collected multimodal streams are usually unaligned. The asynchrony across modalities increases the difficulty on conducting efficient multimodal fusion. Hence, this work mainly focuses on multimodal fusion from unaligned multimodal sequences. To this end, we propose the Progressive Modality Reinforcement (PMR) approach based on the recent advances of crossmodal transformer. Our approach introduces a message hub to exchange information with each modality. The message hub sends common messages to each modality and reinforces their features via crossmodal attention. In turn, it also collects the reinforced features from each modality and uses them to generate a reinforced common message. By repeating the cycle process, the common message and the modalities' features can progressively complement each other. Finally, the reinforced features are used to make predictions for human emotion. Comprehensive experiments on different human multimodal emotion recognition benchmarks clearly demonstrate the superiority of our approach.

1. Introduction

Human multimodal emotion recognition focuses on recognizing the sentiment attitude of humans from video clips [26, 16, 25, 17, 5]. This task involves time-series data of different modalities, e.g., natural language, facial gestures, and acoustic behaviors. The multimodal setting can provide rich information for thorough sentiment understanding. In practice, however, the collected multimodal streams are usually asynchronous, due to the variable sampling rates for sequences from different modalities. For example, the video frame with a depressed facial expression may relate to a negative word spoken in the past. The asynchrony across different modalities can increase the difficulty on conducting efficient multimodal fusion.

The previous works address the above issues by predefined word-level alignment [22, 13, 19, 24]. To this end, the visual and acoustic sequences are first manually aligned in the resolution of the textual words. Multimodal fusion is then conducted on the aligned time steps. However, the manual word-alignment process is usually labor-intensive and requires domain knowledge. Recently, Tsai et al. propose the Multimodal Transformer (MulT) approach to fuse crossmodal information from unaligned data sequences [18]. Their approach introduces the modality reinforcement unit to reinforce a target modality with information from a source modality by learning the directional pairwise attention between elements across modalities (see Fig. 1(a)), based on the recent advances of transformer [20]. By exploring the crossmodal interaction between elements via the crossmodal attention operations, MulT can implement multimodal fusion from asynchronous sequences without explicitly aligning the data.

In their approach, however, the modality reinforcement of each direction is performed independently and does not exchange information with each other. Hence, the multimodal fusion only appears between each directional modality pair, but not across all the modalities involved in human

* Corresponding authors.

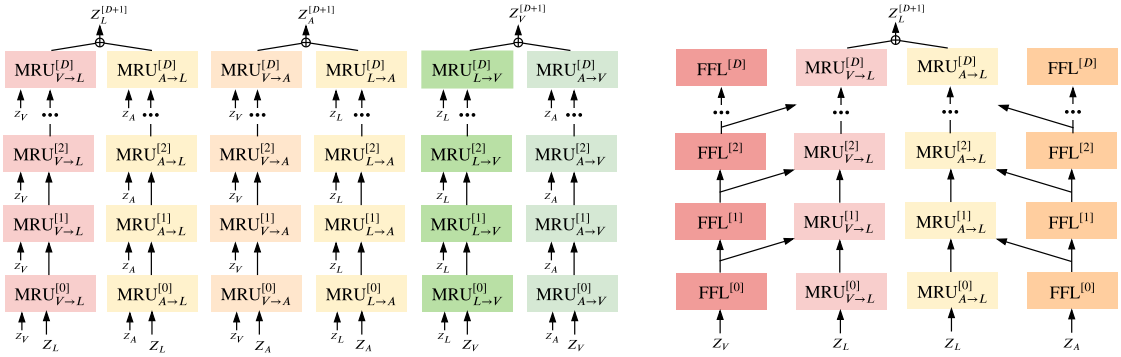

Figure 1: The architecture of the MulT model. $\mathrm{MRU}^{[i]}_{s\to t}$ represents a modality reinforcement unit, in which a source modality s reinforces a target modality t by attending to the crossmodal interaction between elements. $\mathrm{FFL}^{[i]}$ represents a feed-forward layer. (a) the low-level version in which the target modality is reinforced by repeatedly attending to the low-level features of the source modality; (b) the high-level version in which the target modality is reinforced by repeatedly attending to higher-level features of the source modality.

emotion recognition. It is inefficient to fuse the sequences of multiple modalities by using the pairwise manner. For example, the redundant information can be introduced by directly concatenating the visual sequence reinforced by the language modality and that reinforced by the acoustic modality. It is crucial to conduct multimodal fusion by considering the three-way interactions across all the involved modalities.

Moreover, the independent pairwise fusion approach fails to exploit the high-level features of the source modality. For each directional modality pair, as shown in Fig. 1(a), the target modality is reinforced by repeatedly attending to the low-level features of the source modality. Intuitively, the deep interactions cross modalities cannot be explored via the semi-shallow structure. Their approach also notices this problem and attempts to implement crossmodal attention via the high-level features of the source modality by stacking feed-forward layers over the source modality (see Fig. 1(b)). However, reduced performance is observed. This is because that the source branch does not receive clear supervision to update its feed-forward layers, since the modality reinforcement operations mainly focus on generating a reinforced target modality. As a result, it is unclear whether the high-level features of the source modality are better than the low-level features. Instead, the increased modal complexity can reduce the performance.

Motivated by the above observations, this work proposes the Progressive Modality Reinforcement (PMR) approach for multimodal fusion from unaligned multimodal sequences. Our approach introduces a message hub to exchange information with each modality. As shown in Fig. 2, the message hub can send common messages to each modality in order to reinforce their features via crossmodal attention. In turn, it also collects the reinforced features from each modality and uses them to generate an improved common message. In our approach, hence, the common message and the modalities' features progressively complement each other. Moreover, we introduce a dynamic filter mechanism in the modality reinforcement unit to dynamically determine the passed proportions of the reinforced features. Compared with the prior MulT model [18], the advantage of our approach lies in two aspects. First, the common message promotes effective information flow across modalities and encourages the crossmodal attention operations to explore the element-level dependencies across all the three modalities instead of the directional pairwise dependencies. Second, the progressive reinforcement strategy provides an effective way to leverage the high-level features of the source modality for modality reinforcement. Unlike in Fig. 1(b), the feature of the source modality can receive clear supervision in the reinforcement unit where it is considered as the target modality. The superiority of our approach is verified via extensive empirical experiments on different human multimodal emotion recognition benchmarks with both the word-aligned setting and the unaligned setting.

To sum up, the contributions of this work are mainly three-fold:

- We introduce the message hub to explore the three-way interactions across all the involved modalities under the background of multimodal fusion from unaligned multimodal sequences.
- We propose the progressive strategy to leverage the high-level features of the source modality for multimodal fusion.

- Our approach can obtain better results than the existing state-of-the-art works over different human multimodal emotion recognition benchmarks.

2. Related Works

Human multimodal emotion recognition requires to infer the sentiment attitude of humans from video clips [26, 16, 17, 5]. The crucial point lies in multimodal fusion from data sequences of different modalities such as natural language, video frames and acoustic signals [19]. Compared to multimodal fusion from static modalities like images [9, 10, 4, 15], this task requires to fuse crossmodal information from time-series signals. The early works simply adopt the early-fusion strategy by concatenating the input sequences from different modalities [11, 8] or the late-fusion strategy by combining the high-level information learnt from each individual modality [16, 14, 17]. Furthermore, Gan et al. propose to infer the joint representations of different modalities by probabilistic graphical models [5]. Although these prior works obtain better performance than learning from a single modality, they do not explicitly consider the inherent dependencies between elements of sequences from different modalities, which are crucial for efficient multimodal fusion. To this end, the recent works include a manual step to align the visual and acoustic sequences in the resolution of textual words before training [19, 21, 13]. These works perform multimodal fusion on the word-aligned time steps by hierarchical attention mechanism [7], nonverbal temporal interaction [21], cyclic translation [13], etc. However, the manual word-alignment process is usually labor-intensive and time-consuming. Moreover, the word-level multimodal fusion ignores the long-range dependencies between elements from different modalities.

To fuse information from unaligned multimodal sequences, the early work explores the dependencies between elements across modalities according to the maximum mutual information criterion [25]. However, its performance is far from satisfactory due to the shallow learning architecture. Recently, Tsail et al. propose the crossmodal attention mechanism to learn the inherent correlations across modalities [18]. Their approach repeatedly reinforces one modality with information from the other modalities through learning the directional pairwise attention between elements of different modalities.

3. Progressive Modality Reinforcement

3.1. Problem statement

In this work, the human emotion recognition task involves three major modalities, i.e., language (L), video (V), and audio (A). Denote by $X_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d_{\{L,V,A\}}}$ the input sequences from the corresponding modalities. $T_{(.)}$ and $d_{(.)}$ represent the sequence length and feature dimension, respectively. Our goal is to perform efficient multimodal fusion from unaligned multimodal data sequences, in order to obtain the representation that can produce desirable performance in sentiment attitude prediction.

3.2. Preliminary - crossmodal attention.

The crossmodal attention operation reinforces the target modality with information from a source modality by learning the directional pairwise attention between them [18]. Denote by $X_s \in \mathbb{R}^{T_s \times d_s}$ the data sequence from the source modality and $X_t \in \mathbb{R}^{T_t \times d_t}$ the data sequence from the target modality, where $s, t \in \{L, V, A\}$. Similar to the self-attention mechanism, the crossmodal attention unit involves Querys, Keys, and Values, which are defined as $Q_t = X_t W_{Q_t}$ with $W_{Q_t} \in \mathbb{R}^{d_t \times d_k}$, $K_s = X_s W_{K_s}$ with $W_{K_s} \in \mathbb{R}^{d_s \times d_k}$, and $V_s = X_s W_{V_s}$ with $W_{V_s} \in \mathbb{R}^{d_s \times d_v}$, respectively. One individual head of crossmodal attention is defined as:

$$\begin{aligned} Y_t &= \mathrm{CA}_{s \to t}(X_s, X_t) \\ &= \mathrm{softmax}(\frac{Q_t K_s^T}{\sqrt{d_k}}) V_s \\ &= \mathrm{softmax}(\frac{X_t W_{Q_t} W_{K_s}^T X_s^T}{\sqrt{d_k}}) X_s W_{V_s}, \end{aligned}$$

where $Y_t \in \mathbb{R}^{T_t \times d_v}$. The full crossmodal attention operation with h heads is represented as $Y_t = \mathrm{CA}^{\mathrm{mul}}_{s \to t}(X_s, X_t)$, where $Y_t \in \mathbb{R}^{T_t \times h d_v}$. The target modality is reinforced by encouraging the model to attend to crossmodal interaction between elements.

3.3. Model overview

Our model is trained in an end-to-end manner. Following [18], we use a 1D temporal convolutional layer to process the input sequences and then augment them by the positional embedding. Denote by $Z_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d}$ the processed sequences. Note that the 1D temporal convolutional layer projects the features of different modalities to the identical dimension by controlling the kernel size used for each modality. The common message is initialized by concatenating the low-level sequence from each modality: $Z_C = [Z_L, Z_V, Z_A]$, where $Z_C \in \mathbb{R}^{T_C \times d}$ and $T_C = T_L + L_V + L_A$. Immediately, the modality reinforcement layers will repeatedly reinforce Z_C and $Z_{\{L,V,A\}}$ by exploiting the correlations between elements across modalities. Fig. 2 displays the information flow across the modality reinforcement layers. We then concatenate the reinforced features as $[Z_C, Z_L, Z_V, Z_A] \in \mathbb{R}^{2T_C \times d}$ and pass it through a transformer layer. Finally, several fully-connected layers are included to make the predictions for human emotion.

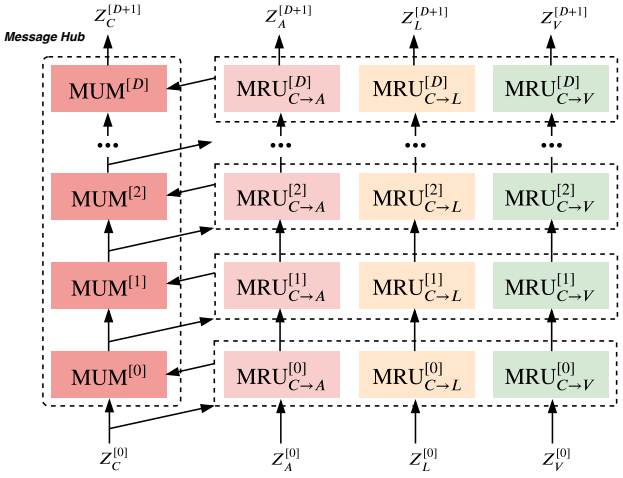

Figure 2: The information flow across the modality reinforcement layers of the proposed model. $\mathrm{MUM}^{[i]}$ represents the message update module in which the common message is reinforced by the reinforced modalities' features from the next layer. $\mathrm{MRU}^{[i]}_{C\to *}$ represents the modality reinforcement unit in which the corresponding target modality is reinforced by the common message.

3.4. Progressive modality reinforcement

Initially, both the modalities' features $Z_{\{L,V,A\}}$ and the common message Z_C do not carry information about the interactive relationship between different modalities, which is crucial for efficient multimodal fusion. In the modality reinforcement layers, Z_C and $Z_{\{L,V,A\}}$ progressively complement each other by exploiting the inherent correlations between elements across modalities. To be specific, each layer includes three modality reinforcement units for updating the modalities' features $Z_{\{L,V,A\}}$ and one message update module for updating the common message Z_C. Denote by $\mathrm{MUM}^{[i]}$ the message update module and $\mathrm{MRU}^{[i]}_{C\to *}$ the modality reinforcement unit for the corresponding modality, where $* \in \{L, V, A\}$. The superscript $[i]$ indicates the i-th modality reinforcement layer.

Modality reinforcement unit. The architecture of the modality reinforcement unit $\mathrm{MRU}^{[i]}_{C\to *}$ is shown in Fig. 3(a). It takes $Z^{[i]}_C$ and $Z^{[i]}_*$ as its inputs and outputs the reinforced features $Z^{[i+1]}_*$:

$$Z^{[i+1]}_* = \mathrm{MRU}^{[i]}_{C\to *}(Z^{[i]}_C, Z^{[i]}_*),$$

where $* \in \{L, V, A\}$ and $Z^{[i+1]}_* \in \mathbb{R}^{T_*\times d}$. Unlike in MulT, all the three modalities will participate in a single modality reinforcement unit via the common message.

Specifically, $\mathrm{MRU}^{[i]}_{C\to *}$ reinforces $Z^{[i]}_*$ by two branches, including a self-attention one and a crossmodal attention one:

$$Z^{[i]}_{C\to *} = \mathrm{CA}^{\mathrm{mul}}_{C\to *}(\mathrm{LN}(Z^{[i]}_C), \mathrm{LN}(Z^{[i]}_*)),$$
$$Z^{[i]}_* = \mathrm{SA}^{\mathrm{mul}}(\mathrm{LN}(Z^{[i]}_*)),$$

where $Z^{[i]}_{C\to *}, Z^{[i]}_* \in \mathbb{R}^{T_*\times d}$, $\mathrm{CA}^{\mathrm{mul}}$ and LN represent the multi-head self-attention operation and the layer normalization operation, respectively. Immediately, the reinforced features $Z^{[i]}_*$ and $Z^{[i]}_{C\to *}$ are processed via the following dynamic filter mechanism:

$$G^{[i]}_* = \mathrm{sigmoid}(Z^{[i]}_* \cdot W^{[i]}_* + Z^{[i]}_{C\to *} \cdot W^{[i]}_{C\to *} + b^{[i]}_*),$$
$$Z^{[i]}_* = G^{[i]}_* \odot Z^{[i]}_* + (1 - G^{[i]}_*) \odot Z^{[i]}_{C\to *},$$

where $W^{[i]}_* \in \mathbb{R}^{d\times d}$, $W^{[i]}_{C\to *} \in \mathbb{R}^{d\times d}$, and $b^{[i]}_* \in \mathbb{R}^{T_*\times d}$. The passed proportions of each branch can be dynamically determined via the learnable parameters $W^{[i]}_*$ and $b^{[i]}_*$. This operation enables to filter information produced by incorrect crossmodal interactions. Finally, as in the transformer model [20], a position-wise feed-forward layer with skip connection will process $Z^{[i]}_*$ and generate $Z^{[i+1]}_*$ for the next modality reinforcement layer.

Message update module. The reinforced features $Z^{[i+1]}_{\{L,V,A\}}$ will also be used to reinforce the common message $Z^{[i]}_C$ in the previous modality reinforcement layer. The architecture of $\mathrm{MUM}^{[i]}$ is displayed in Fig. 3(b). It takes $Z^{[i]}_C$ and $Z^{[i+1]}_{\{V,L,A\}}$ as its inputs and outputs the reinforced common message $Z^{[i+1]}_C$:

$$Z^{[i+1]}_C = \mathrm{MUM}^{[i]}(Z^{[i+1]}_V, Z^{[i+1]}_L, Z^{[i+1]}_A, Z^{[i]}_C).$$

Specifically, the message update module includes three modality reinforcement units, each of which reinforces the common message $Z^{[i]}_C$ by one modality. Denote by $\mathrm{MRU}^{[i]}_{*\to C}$ the corresponding modality reinforcement unit, where $* \in \{V, L, A\}$. In $\mathrm{MRU}^{[i]}_{*\to C}$, $Z^{[i]}_C$ is reinforced by attending to the elements of $Z^{[i+1]}_*$:

$$Z^{[i]}_{*\to C} = \mathrm{MRU}^{[i]}_{*\to C}(Z^{[i+1]}_*, Z^{[i]}_C),$$

where $Z^{[i]}_{*\to C} \in \mathbb{R}^{T_C\times d}$. The three-way interactions across all the involved modalities can be explored via the self-attention operation over $Z^{[i]}_C$ (see Fig.3(a)). Immediately, $Z^{[i]}_{*\to C}$ is fused into $Z^{[i]}_C$ via an attention layer. To this end, we first obtain a reshaped counterpart of $Z^{[i]}_{*\to C}$: $\hat{Z}^{[i]}_{*\to C} \in \mathbb{R}^{T_C\cdot d\times 1}$ and then process them as follows:

$$\mu^{[i]}_{*\to C} = U^T \tanh(W^{[i]}_{*\to C} \cdot \hat{Z}^{[i]}_{*\to C} + b^{[i]}_{*\to C}),$$
$$\alpha^{[i]}_{*\to C} = \frac{\exp(\mu^{[i]}_{*\to C})}{\sum_{*\in\{L,V,A\}} \exp(\mu^{[i]}_{*\to C})},$$
$$Z^{[i]}_C = \sum_{*\in\{V,L,A\}} \alpha^{[i]}_{*\to C} \odot Z^{[i]}_{*\to C},$$

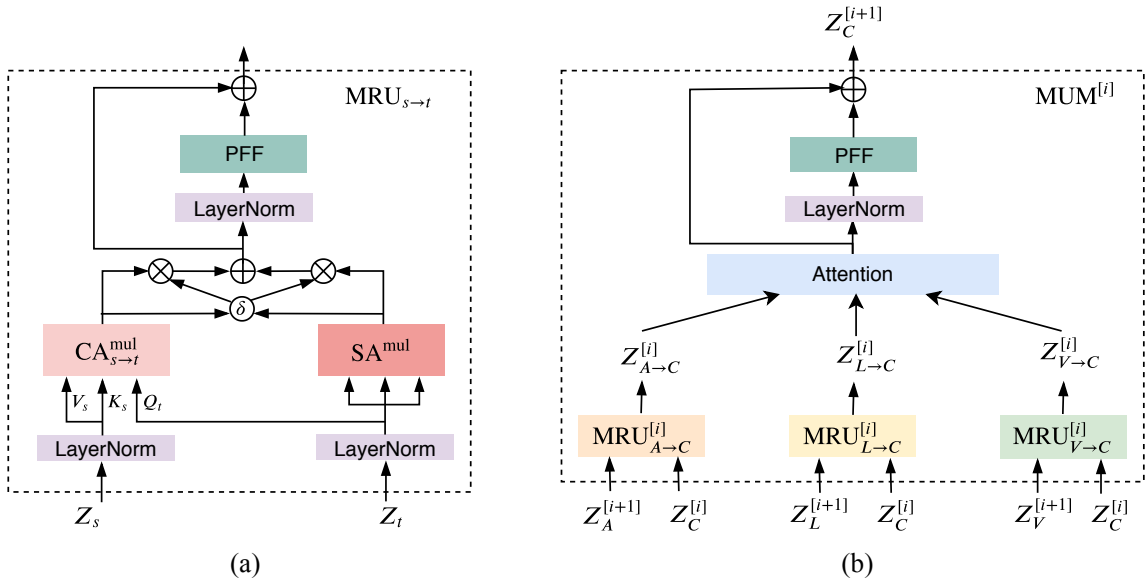

Figure 3: (a) The modality reinforcement unit in which the features of a target modality are reinforced by a source modality via crossmodal attention. PFF represents the positionwise feed-forward layer. $\text{CA}^{\text{mul}}_{s\to t}$ represents the crossmodal attention operation. SA^{mul} represents the self-attention operation. (b) The message update module in which the common message is updated by the reinforced features of each modality.

where $U \in \mathbb{R}^{T_C\cdot d\times 1}$, $W^{[i]}_{*\to C} \in \mathbb{R}^{T_C\cdot d\times T_C\cdot d}$ and $b^{[i]}_{*\to C} \in \mathbb{R}^{T_C\cdot d\times 1}$ are learnable parameters. The attention layer can dynamically control the passed information in $Z^{[i]}_{*\to C}$ and generate an informative common message. Finally, we pass $Z_C^{[i]}$ through a position-wise feed-forward layer with skip connection and obtain the output $Z_C^{[i+1]}$.

In the next modality reinforcement layer, $Z_C^{[i+1]}$ will again reinforce $Z_*^{[i+1]}$ via the $\text{MRU}^{[i+1]}_{C\to *}$ unit. Compared with the prior MulT model [18] which reinforces the target modality by repeatedly attending to the low-level features of the source modality, our approach enables Z_* and Z_C to progressively complement each other, i.e., Z_C reinforces Z_*, and in turn, the reinforced Z_* reinforces Z_C to produce a better common message. We also note that our approach will contain fewer modality reinforcement units at each layer if more modalities are involved (i.e., A_n^2 for MulT and $2n$ for our approach with n indicating the modality number). Algorithm 1 displays the information flow across the modality reinforcement layers.

Algorithm 1 The forward propagation procedure of the modality reinforcement layers.

Input: the sequences processed by 1D temporal convolution and positional embedding: $Z_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}}\times d}$; the layer number: D.

Output: the reinforced modalities' features: $Z_{\{L,V,A\}}$; the reinforced common message: Z_C.

Initialize the modalities' features: $Z^{[0]}_{\{L,V,A\}} = Z_{\{L,V,A\}}$;
Initialize the common message: $Z_C^{[0]} = [Z_L, Z_V, Z_A]$;
$i = 0$;
while $i \leqslant D$ **do**
  Update the modalities' features: $Z_*^{[i+1]} = \text{MRU}^{[i]}_{C\to *}(Z_C^{[i]}, Z_*^{[i]})$, where $* \in \{L, V, A\}$;
  Update the common message: $Z_C^{[i+1]} = \text{MUM}^{[i]}(Z_V^{[i+1]}, Z_L^{[i+1]}, Z_A^{[i+1]}, Z_C^{[i]})$;
  $i = i + 1$;
end while
$Z_C = Z_C^{[D+1]}$; $Z_{\{L,V,A\}} = Z^{[D+1]}_{\{L,V,A\}}$.
return Z_C and $Z_{\{L,V,A\}}$.

4. Experiments

4.1. Experimental setup

We follow the common protocol of the prior works [18, 19, 21] and conduct experiments on the standard human multimodal emotion recognition benchmarks, including CMU-MOSI [24], CMU-MOSEI [23] and IEMOCAP [2]. The experiments are conducted on both the word-aligned and unaligned settings.

CMU-MOSI is a dataset that contains 2,199 samples of

Table 1: The hyperparameter settings adopted in each human multimodal emotion recognition benchmark.

Setting	CMU-MOSEI	CMU-MOSI	IEMOCAP
Optimizer	Adam	Adam	Adam
Batch size	32	64	32
Learning rate	1e-3	1e-3	1e-3
Epoch number	100	120	60
Feature size d	40	40	40
Attention head h	10	8	8
Kernel size (L/V/A)	3/3/3	3/3/3	3/3/5
Reinforcement layer D	5	4	4

Table 2: Comparison on the CMU-MOSI benchmark under both the word-aligned setting and the unaligned setting.

Setting	Method	$\mathrm{Acc}_7(\%)$	$\mathrm{Acc}_2(\%)$	F1(%)
Aligned	EF-LSTM	33.7	75.3	75.2
	LF-LSTM	35.3	76.8	76.7
	MFM [19]	36.2	78.1	78.1
	RAVEN [21]	33.2	78.0	76.6
	MCTN [13]	35.6	79.3	79.1
	MulT [18]	40.0	83.0	82.8
	PMR(ours)	**40.6**	**83.6**	**83.4**
Unaligned	EF-LSTM	31.0	73.6	74.5
	LF-LSTM	33.7	77.6	77.8
	RAVEN [21]	31.7	72.7	73.1
	MCTN [13]	32.7	75.9	76.4
	MulT [18]	39.1	81.1	81.0
	PMR(ours)	**40.6**	**82.4**	**82.1**

short monologue video clips [24]. Its predetermined data split includes 1,284 training samples, 229 validation samples and 686 testing samples. The acoustic features and the visual features are extracted at the sampling rate of 12.5 and 15 Hz, respectively. Each multimodal sample has a sentiment score which ranges from -3 (strongly negative) to 3 (strongly positive). In agreement with the prior works [18, 19], we evaluate the performance by the following metrics: 7-class accuracy (i.e., Acc_7), binary accuracy (i.e., Acc_2) and F1 score.

CMU-MOSEI is a dataset that contains 22,856 samples of movie review video clips from YouTube [23]. Its predetermined data split includes 16,326 training samples, 1,871 validation samples and 4,659 testing samples. The acoustic features and the visual features are extracted at the sampling rate of 20 and 15 Hz, respectively. Likewise, each multimodal sample has a sentiment score ranging from -3 to 3. The same performance metrics are employed as in the above setting.

Table 3: Comparison on the CMU-MOSEI benchmark under both the word-aligned setting and the unaligned setting.

Setting	Method	$\mathrm{Acc}_7(\%)$	$\mathrm{Acc}_2(\%)$	F1(%)
Aligned	EF-LSTM	47.4	78.2	77.9
	LF-LSTM	48.8	80.6	80.6
	G-MFN [24]	45.0	76.9	77.0
	RAVEN [21]	50.0	79.1	79.5
	MCTN [13]	49.6	79.8	80.6
	MulT [18]	51.8	82.5	82.3
	PMR(ours)	**52.5**	**83.3**	**82.6**
Unaligned	EF-LSTM	46.3	76.1	75.9
	LF-LSTM	48.8	77.5	78.2
	RAVEN [21]	45.5	75.4	75.7
	MCTN [13]	48.2	79.3	79.7
	MulT [18]	50.7	81.6	81.6
	PMR(ours)	**51.8**	**83.1**	**82.8**

IEMOCAP is a dataset that contains 4,453 samples of video clips [2]. Its predetermined data split includes 2,717 training samples, 798 validation samples and 938 testing samples. The acoustic and visual features are extracted at the sampling rate of 12.5 and 15 Hz, respectively. Following [21], we focus on recognizing 4 kinds of emotions (i.e., happy, sad, angry and neutral) in each video clip. Moreover, this setting is established as a multi-label task, since the sad and the angry emotions can exit in a video clip simultaneously. In agreement with the prior works [19, 21], we evaluate the performance by the binary classification accuracy and the F1 score for each emotion class.

4.2. Implementation details

To extract features of the textual modality, we convert the video transcripts into the pre-trained Glove model to obtain 300-dimensional word embeddings [12]. For the visual modality, we process the video frames by Facet to generate 35 facial action units that represent the facial muscle movement [1]. To extract features of the acoustic modality, we process the acoustic signals by COVAREP to obtain 74-dimensional features [3].

Table 1 displays the hyperparameters used in each benchmark. The kernel size relates to the 1D temporal convolutional layer which is used to process the input sequences. In each benchmark, both the crossmodal attention operation and the self-attention operation use the same number of attention heads. The hyper-parameters are determined on the validation set.

Table 4: Comparison on the IEMOCAP benchmark under both the word-aligned setting and the unaligned setting. The performance is evaluated by the binary classification accuracy and the F1 score for each emotion class.

Setting	Method	Happy		Sad		Angry		Neutral	
		Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
Aligned	EF-LSTM	86.0	84.2	80.2	80.5	85.2	84.5	67.8	67.1
	LF-LSTM	85.1	86.3	78.9	81.7	84.7	83.0	67.1	67.6
	MFM [19]	90.2	85.8	88.4	86.1	87.5	86.7	72.1	68.1
	RAVEN [21]	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3
	MCTN [13]	84.9	83.1	80.5	79.6	79.7	80.4	62.3	57.0
	MulT [18]	90.7	88.6	86.7	86.0	87.4	87.0	72.4	70.7
	PMR(ours)	**91.3**	**89.2**	**87.8**	**87.0**	**88.1**	**87.5**	**73.0**	**71.5**
Unaligned	EF-LSTM	76.2	75.7	70.2	70.5	72.7	67.1	58.1	57.4
	LF-LSTM	72.5	71.8	72.9	70.4	68.6	67.9	59.6	56.2
	RAVEN [21]	77.0	76.8	67.6	65.6	65.0	64.1	62.0	59.5
	MCTN [13]	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3
	MulT [18]	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
	PMR(ours)	**86.4**	**83.3**	**78.5**	**75.3**	**75.0**	**71.3**	**63.7**	**60.9**

Table 5: Ablation study on the CMU-MOSEI benchmark under the unaligned setting. In each row, the corresponding component is progressively included into the model.

Model design	$\text{Acc}_7(\%)$	$\text{Acc}_2(\%)$	F1(%)
Low-level feature [18]	50.7	81.6	81.6
High-level feature [18]	50.3	80.5	80.6
+ Progressive strategy	51.2	82.4	82.2
+ Message hub	51.6	82.8	82.6
+ Dynamic filter (full model)	**51.8**	**83.1**	**82.8**

4.3. Performance comparison

The proposed approach is compared to the existing state-of-the-art baselines, including Early Fusion LSTM (EF-LSTM), Late Fusion LSTM (LF-LSTM), Multimodal Factorization Model (MFM) [19], Graph-MFN (G-MFN) [24], Recurrent Attended Variation Embedding Network (RAVEN) [21], Multimodal Cyclic Translation Network (MCTN) [13], Multimodal Transformer (MulT) [18]. Of these, MulT and LF-LSTM can be applied directly to the unaligned setting. For the other methods, we include the Connectionist Temporal Classification (CTC) alignment loss [6] into the learning objective, in order to make them suitable for the unaligned setting.

Word-aligned setting. This setting requires an extra step to manually align the visual and acoustic streams in the resolution of textual words. The multimodal fusion is then conducted on the word-aligned time steps. We display the experimental results of each approach in the upper part of Table 2 - 4. Compared with the other baselines, our proposed approach obtains better performance on different metrics over all the three benchmarks.

Unaligned setting. This setting requires to fuse crossmodal information directly from the unaligned multimodal sequences and is more challenging than the word-aligned setting. We display the comparison of each approach in the bottom part of Table 2 - 4. We can draw the following observations. First, except for MulT, most of the compared baselines obtain poor performance on the unaligned setting since their models do not consider the crossmodal interaction between elements. Second, our approach can outperform MulT on different metrics over all the three benchmarks. We can see that the performance improvement of our approach is more significant in the unaligned setting than in the word-aligned setting. This observation indicates that the technical superiority of our approach mainly lies in better capturing the dependencies between elements across modalities, which is consistent with our motivation.

4.4. Analysis

Ablation study. Table 5 displays the ablation study on the CMU-MOSEI benchmark. The first two rows display the performance of the MulT model implemented by the low-level version (see Fig. 1(a)) and the high-level version (see Fig. 1(b)), respectively. The high-level version of MulT is implemented by stacking feed-forward layers over the source modality. We can see that worse performance is obtained by the high-level version, which seems unreasonable at first sight. This can be attributed to the reason that the source modality of each directional modality pair cannot re-

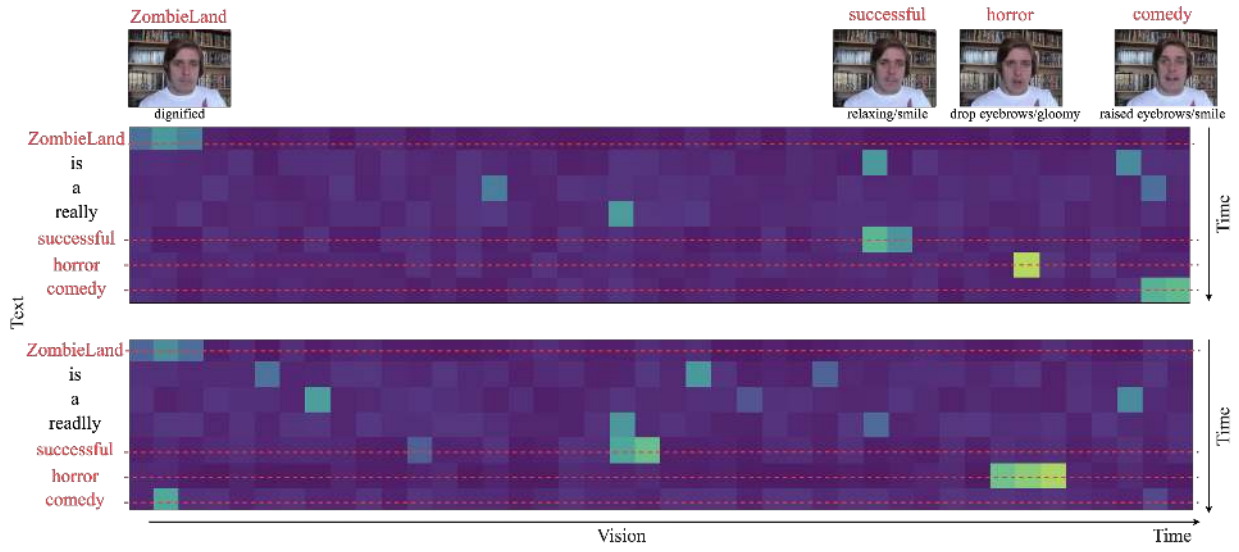

Figure 4: Visualization for the crossmodal correlations on the CMU-MOSI benchmark. The visualization samples of our approach and MulT are displayed in the upper part and the bottom part, respectively. We conduct the visualization by observing the crossmodal attention weights of the corresponding modality reinforcement unit (i.e., $\mathrm{MRU}_{C\to L}$ for the proposed approach and $\mathrm{MRU}_{V\to L}$ for MulT) in the fourth reinforcement layer. The textual words which are closely related to human emotion recognition are displayed in red. The textual words above the video clips are the corresponding spoken words.

ceive clear supervision to be updated due to the independent reinforcement structure of MulT.

In the next row, we introduce the progressive reinforcement strategy for each modality pair. To this end, the reinforced target modality is in turn used to reinforce the source modality, e.g., the visual modality and the language modality progressively reinforce each other. Unlike in MulT, the source modality can receive clear supervision to be updated since it is a target modality as well. The progressive reinforcement strategy provides an efficient way to leverage the high-level features of the source modality. The performance from the third row clearly verifies the above discussion. Moreover, we introduce the message hub to exchange information with each modality. Improved performance can be observed from the fourth row. This component improves the performance by encouraging the model to explore the three-way interactions across all the modalities. Finally, we introduce the dynamic filter mechanism in the modality reinforcement unit, which can also make an effective contribution to multimodal fusion.

Qualitative analysis. Fig. 4 displays the visualization for the crossmodal interaction between elements. The visualization sample of our approach and MulT are displayed in the upper part and the bottom part, respectively. We can see that our approach can correlate the emotion related textual word with the corresponding video clips well. Compared to MulT, our approach can encourage the model to attend to more meaningful signals across the two modalities. From the visualization sample of MulT, the correlations between the textual words and the video clips are not clear.

5. Conclusion

This work proposes the progressive modality reinforcement approach towards multimodal fusion from unaligned multimodal sequences, under the background of human multimodal emotion recognition. To this end, we introduce a message hub to exchange information with each modality. The message hub can encourage a more efficient multimodal fusion by exploring the inherent correlations across all the modalities via the common message. Moreover, the common message and the modalities' features progressively complement each other by attending to the crossmodal interaction between elements. The progressive reinforcement strategy provides an effective way to leverage the high-level features of the source modality in modality reinforcement. The experimental results over different benchmarks clearly demonstrate that our approach obtains better results than the existing state-of-the-art works.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No.11829101) and the Fundamental Research Funds for the Central Universities of China (No.JBK1806002). G. Lin's participation was supported by an NTU Start-up Grant and MOE Tier-1 research grants: RG28/18 (S), RG22/19 (S) and RG95/20.

References

[1] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *WACV*, pages 1–10, 2016.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359, 2008.

[3] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. COVAREP - A collaborative voice analysis repository for speech technologies. In *ICASSP*, pages 960–964, 2014.

[4] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE Trans. Medical Imaging*, 38(5):1116–1126, 2019.

[5] Quan Gan, Shangfei Wang, Longfei Hao, and Qiang Ji. A multimodal deep regression bayesian network for affective video content analyses. In *ICCV*, pages 5123–5132, 2017.

[6] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In William W. Cohen and Andrew W. Moore, editors, *ICML*, volume 148, pages 369–376, 2006.

[7] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *ACL*, pages 2225–2235, 2018.

[8] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. In *NAACL*, pages 153–163, 2015.

[9] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, pages 10142–10151, 2019.

[10] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *CVPR*, pages 14222–14231, 2020.

[11] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[13] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, pages 6892–6899, 2019.

[14] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. Multimodal emotion recognition using deep learning architectures. In *WACV*, pages 1–9, 2016.

[15] Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi. Misalignment-robust joint filter for cross-modal image pairs. In *ICCV*, pages 3315–3324, 2017.

[16] Dung Nguyen Tien, Kien Nguyen, Sridha Sridharan, David Dean, and Clinton Fookes. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Comput. Vis. Image Underst.*, 174:33–42, 2018.

[17] Dung Nguyen Tien, Kien Nguyen Thanh, Sridha Sridharan, Afsane Ghasemi, David Dean, and Clinton Fookes. Deep spatio-temporal features for multimodal emotion recognition. In *WACV*, pages 1215–1223, 2017.

[18] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569, 2019.

[19] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[21] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, pages 7216–7223, 2019.

[22] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, pages 5634–5641, 2018.

[23] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACL*, pages 2236–2246, 2018.

[24] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.*, 31(6):82–88, 2016.

[25] Zhihong Zeng, Jilin Tu, Brian Pianfetti, Ming Liu, Tong Zhang, ZhenQiu Zhang, Thomas S. Huang, and Stephen E. Levinson. Audio-visual affect recognition through multi-stream fused HMM for HCI. In *CVPR*, pages 967–972, 2005.

[26] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur A. Ciftci, Shaun J. Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPP*, pages 3438–3446, 2016.