QAP: A Quantum-Inspired Adaptive-Priority-Learning Model for Multimodal Emotion Recognition

Ziming Li[1,2], **Yan Zhou**[1,*], **Yaxin Liu**[1,2], **Fuqing Zhu**[1],
Chuanpeng Yang[1,2], **Songlin Hu**[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{liziming, zhouyan, liuyaxin, zhufuqing, yangchuanpeng, husonglin}@iie.ac.cn

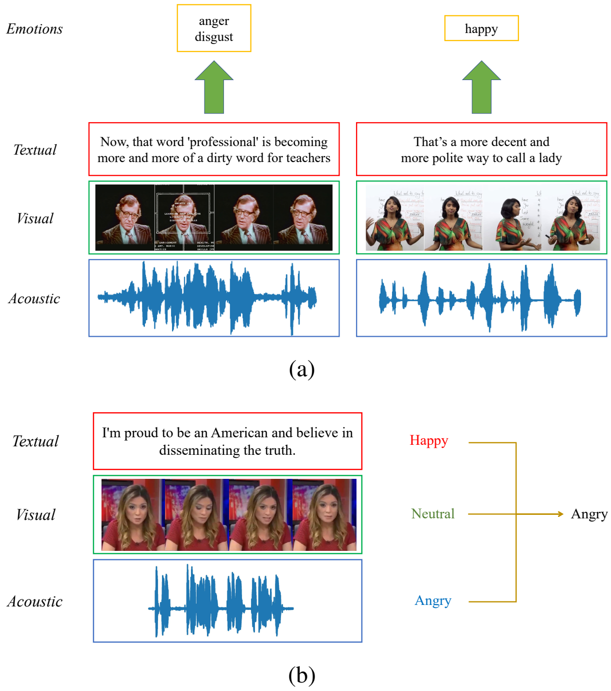
Abstract

Multimodal emotion recognition for video has gained considerable attention in recent years, in which three modalities (*i.e.,* textual, visual and acoustic) are involved. Due to the diverse levels of informational content related to emotion, three modalities typically possess varying degrees of contribution to emotion recognition. More seriously, there might be inconsistencies between the emotion of individual modality and the video. The challenges mentioned above are caused by the inherent uncertainty of emotion. Inspired by the recent advances of quantum theory in modeling uncertainty, we make an initial attempt to design a quantum-inspired adaptive-priority-learning model (QAP) to address the challenges. Specifically, the quantum state is introduced to model modal features, which allows each modality to retain all emotional tendencies until the final classification. Additionally, we design Q-attention to orderly integrate three modalities, and then QAP learns modal priority adaptively so that modalities can provide different amounts of information based on priority. Experimental results on the IEMOCAP and MOSEI datasets show that QAP establishes new state-of-the-art results.

1 Introduction

Multimodal emotion recognition (MER) has attracted more and more interest due to the rapid growth of multimedia information. MER aims to recognize the emotions of the speaker in the video. Multiple modalities enrich human emotional expression and they are all closely related to emotion. Generally, textual modality provides the most basic semantic information, visual modality provides emotional expressions, and acoustic modality provides the changing tone.

Three modalities also bring greater challenges to emotion recognition. Due to the different amounts of information related to emotions, the priority of each modality varies from sample to sample. If different modalities are not discriminated in fusion, information related to emotion cannot be fully extracted. In the example on the left of Figure 1 (a), the dejected expression, wrinkled eyebrows and drooping corners of the eyes all show anger and disgust, so visual modality contributes more to emotion. In the example on the right, a rising tone shows the emotion of happiness, so acoustic modality has higher priority than visual modality. Most previous works (Tsai et al., 2019; Akhtar et al., 2019; Chauhan et al., 2020) treat modalities equally and do not pay attention to the important role of modal priority. Some other works (Li et al., 2022) integrate modalities in a certain order, but the order is not adaptively adjusted for different samples. In practical scenarios, a fixed order cannot fit all samples.


Emotions
anger disgust
happy
Textual
Now, that word 'professional' is becoming more and more of a dirty word for teachers
That's a more decent and more polite way to call a lady
Visual
Acoustic
Textual
I'm proud to be an American and believe in disseminating the truth.
Visual
Acoustic
Happy
Neutral
Angry
Angry

(a)

(b)

Figure 1: Examples of multimodal emotion recognition.

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 12191–12204
July 9-14, 2023 ©2023 Association for Computational Linguistics

More seriously, there might be inconsistencies between the emotion of individual modality and the video. In the example in Figure 1 (b), happiness is expressed in the text modality, but the emotion of the video is anger. Some previous methods (Sun et al., 2022; Yang et al., 2022; Yuan et al., 2021) do not consider this issue and still integrate three modalities together, resulting in a negative impact on final emotions. Some other methods (Mittal et al., 2019) remove the modality with inconsistent emotions and replace it with other features, which lose the semantic information contained in the modality.

As part of human cognition, emotion is always in an uncertain state and constantly evolving until the final decision is made (Busemeyer and Bruza, 2012). Specifically, the emotion of a video is considered to be uncertain until it is measured and collapses to an eigenstate, and so does one of the modalities. Conceptually, in non-quantum models, the emotions in the video are pre-defined values and a measurement (classification) merely records them. In other words, three modalities are always aligned to a certain emotional label throughout the entire process before recognition. However, the generation of emotions is often spontaneous and intuitive so the cognitive system is fundamentally uncertain and in an indefinite state. In quantum-like frameworks, the emotion of each modality is treated as an indefinite state (Busemeyer and Bruza, 2012). The final quantum measurement creates a definite state and changes the state of the system. Despite the above advantages compared to previous models, it is also challenging to complete the MER task in a quantum-like framework due to the complex processes such as feature extraction and modal fusion. Technically, we must ensure that the model conforms to the evolution process of the quantum system and that the characteristics of the density matrix remain unchanged.

Inspired by the excellent performance of quantum-like networks in other tasks (Jiang et al., 2020; Liu et al., 2021; Li et al., 2019, 2021), we propose an adaptive-priority-learning model (**QAP**) for MER. QAP uses quantum states instead of traditional feature vectors to represent modal features from the initial feature extraction step to the final emotion classification step. Modal features in each step no longer correspond solely to the final emotional label, but are in a state where emotion is uncertain. In this way, the opposite emotion of a single modality will not affect the final emotion because all modalities are in an uncertain state with multiple emotions.

In MER, it is an inherent problem to effectively extract the features of the raw modalities. Previous works either use pre-extracted features with handcrafted algorithms or extract end-to-end features with pre-trained models. But these two features are not effectively combined together. In QAP, the complex-valued density matrix is used as the unit of modal representation due to the stronger representation ability (Balkır, 2014). By this means, end-to-end features and pre-extracted features are effectively combined by a non-linear method.

For the fusion in the quantum-like framework, $\mathrm{Q\text{-}attention}$ based on the density matrix is designed to orderly integrate the three modalities. After that, since three modalities can form several fusion orders, we use a quantum measurement operator to select the most appropriate fusion order. In this way, QAP can learn modal priority adaptively. Finally, we use another quantum measurement operator to collapse all states in the density matrix to the pure state representing emotion to recognize the emotion.

The main contributions of our paper are as follows:

- We propose QAP, a quantum-inspired adaptive-priority-learning model for multimodal emotion recognition, where each modality is in a state where emotion is uncertain. So modalities with different emotions can be integrated.
- QAP utilizes the density matrix to represent modal features and two kinds of features can be combined effectively. Based on the density matrix, we design $\mathrm{Q\text{-}attention}$ to integrate modalities in order of priority and utilize a quantum measurement operator to select fusion order. So QAP can adaptively learn the modal priority.
- Experimental results on the IEMOCAP (Busso et al., 2008) and CMU-MOSEI (Zadeh and Pu, 2018) datasets show the state-of-the-art performance of QAP.

2 Related Work

MER has attracted more and more attention, and many methods have been used to integrate modalities. Direct concatenation and outer product (Zadeh et al., 2017) are used as fusion methods in the early years. And then Zadeh et al. (2018) proposes a method based on recurrent neural network and designs a gate mechanism. In recent years, models based on attention mechanism (Vaswani et al., 2017; Tsai et al., 2019) are applied to MER and followed by later works. Rahman et al. (2020) proposes an attachment to enable pre-trained models to integrate multimodal information. Zhang et al. (2020) models the dependencies between labels and between each label and modalities for multi-label MER. Hu et al. (2022) presents a graph-based network to capture multimodal features and contextual dependencies. These works treat modalities equally and do not pay attention to modal priority. Li et al. (2022) integrates three modalities in a certain order, but cannot adaptively learn modal priority. In addition, the end-to-end models (Dai et al., 2021; Wei et al., 2022; Wu et al., 2022) are also proposed to make better use of the raw modal information. However, they introduce noise irrelevant to emotion and also ignore the importance of modal priority. The issues of inconsistent emotions and differentiated contributions have not been resolved in the above work, which negatively affects the performance of the model. In contrast, our approach can adaptively learn modal priority and modalities with more emotional information will make a greater contribution

Quantum-inspired or quantum-like models have a good performance in different tasks. Sordoni et al. (2013) first applies the quantum-like model to the field of information retrieval. Li et al. (2019) and Zhang et al. (2018) design the quantum language models in the text matching task. Li and Hou (2021) combines the quantum-like model and the convolutional neural network, and gets an expected result in the sentiment analysis task. Gkoumas et al. (2021b) proposes the first quantum-like model for multimodal sentiment analysis, which is a decision-level fusion framework. Liu et al. (2021) uses quantum interference to integrate textual modality and visual modality. Gkoumas et al. (2021a) introduces the concept of quantum entanglement to multimodal fusion and Li et al. (2021) designs a quantum-like recurrent neural network to model context information. All these works prove that quantum-inspired networks have advantages in modeling human cognitive uncertainty. However, the modules of modal fusion in them are too simple to fully capture the inter-modality information. Besides, integrating three modalities in a quantum-like framework is a challenging task, and we make an initial attempt in this field to make the modalities with opposite emotions be integrated effectively.

3 Preliminaries on Quantum Theory

The construction of a quantum-inspired model is based on quantum theory (QT) (Fell et al., 2019; Busemeyer and Bruza, 2012). In this section, we will briefly introduce the basic concepts of QT. The state vector in QT is defined on a Hilbert space $\mathbb{H}$, which is an infinite inner product space over the complex field. With Dirac Notations, we denote a complex unit vector $\vec{\mu}$ as a ket $|u\rangle$, and its conjugate transpose $\vec{\mu}^H$ is denoted as a bra $\langle u|$. The inner product and outer product of two state vectors $|u\rangle$ and $|v\rangle$ are denoted as $\langle u|v\rangle$ and $|u\rangle \langle v|$.

3.1 State

A quantum state $|\psi\rangle$ is a complete description of a physical system and is a linear superposition of an orthonormal basis in the Hilbert space. The state of a system composed of a single particle is called a pure state. The mathematical form of $|\psi\rangle$ is a complex column vector.

A pure state can also be expressed as a density matrix: $\rho = |\psi\rangle \langle\psi|$. When several pure states are mixed together in the way of classical probability, we use the mixed state to describe the system. The density matrix can also represent a mixed state: $\rho = \sum_{i=1}^{n} p_i |\psi_i\rangle\langle\psi_i|$, where p_i denotes the probability distribution of each pure state and $\sum_{i=1}^{n} p_i = 1$.

In MER, one modality is composed of several tokens, and each token can be regarded as a particle. Therefore, we use the density matrix to represent the modal features which can be viewed as mixed states.

3.2 Evolution

In QT, a state does not remain unchanged, but can evolve over time. The evolution is described by a unitary operator U. U is a complex unitary matrix satisfying $UU^H = I^2$. The evolution process is as follows:

$$\rho^{'} = U\rho U^H, \tag{1}$$

It can be proved that $\rho^{'}$ is also a density matrix as long as ρ is a density matrix. We draw an analogy between the evolution process and the linear transformation process of a density matrix.

3.3 Measurement

Quantum measurement causes a pure state to collapse to a base with a probability. The measurement process is described by an observable M:

$$M = \sum_{j=1}^{n} \lambda_j \left| m_j \right\rangle \left\langle m_j \right|, \tag{2}$$

where $\{|m_j\rangle\}$ are the eigenstates of the operator and also form an orthonormal basis in the Hilbert space. $\{\lambda_j\}$ are the eigenvalues corresponding to eigenstates. According to the Born's rule (Halmos, 2017), the probability of the pure state $|\psi\rangle$ to collapse onto the basis state $|m_j\rangle$ is calculated as follows:

$$p_j = |\left\langle m_j \right| \psi\rangle|^2 = tr(\rho \left| m_j \right\rangle \left\langle m_j \right|), \tag{3}$$

where $\rho = |\psi\rangle\langle\psi|$. For a mixed state, the probability of collapsing to an eigenstate is the weighted sum of the probability values of all pure states. We exploit quantum measurement to calculate the weight of different fusion orders and recognize the final emotions.

4 Model

In this section, we will describe the details of QAP. The overall architecture of QAP is shown in Figure 2. QAP consists of three modules: *Unimodal Complex-valued Representation*, *Adaptive Priority Learning* and *Emotion Recognition*. Firstly, the complex density matrix of the single modality is constructed for modal representation(Section 4.1). In the representation, end-to-end features and pre-extracted features are respectively aligned to the amplitude and the phase of the complex value. Secondly, Q-attention is designed to integrate three modalities orderly and we use a quantum measurement operator to select the appropriate order, then QAP can learn modal priority adaptively (Section 4.2). Finally, another measurement operator is employed to recognize the final emotion (4.3).

4.1 Unimodal Complex-valued Representation

Early works (Zadeh et al., 2018; Zeng et al., 2021) usually extract features with hand-crafted algorithms, but these pre-extracted features cannot be further fine-tuned on different tasks and have poor generalization. In recent years, some methods (Dai et al., 2021; Wei et al., 2022) utilize pre-trained models to extract more modal information, which can be fine-tuned on different tasks. However, fully end-to-end models may bring noise, such as the part outside the face in the image. These noises will cause semantic drift and affect the judgment of video emotion.

To alleviate this problem, we utilize the two kinds of modal features together with complex-valued representation. A complex value can be expressed in polar form: $z = re^{i\theta}$, where r is the amplitude and θ is the phase or argument. So a pure state can be expressed as:

$$\begin{aligned} |\psi\rangle &= [r_1 e^{i\theta_1}, r_2 e^{i\theta_2}, \ldots, r_n e^{i\theta_n}]^T \\ &= [r_1, r_2, \ldots, r_n]^T \odot e^{i[\theta_1, \theta_2, \ldots, \theta_n]^T}, \end{aligned} \tag{4}$$

where $\odot$ is the element-wise product. By formula (4), a pure state can be decomposed from a complex vector into two real vectors: $\vec{r} = [r_1, r_2, \ldots, r_n]^T$ and $\vec{\theta} = [\theta_1, \theta_2, \ldots, \theta_n]^T$. So we just need to construct these two real vectors. On the whole, the end-to-end feature is used as $\vec{r}$, and the pre-extracted feature is used as $\vec{\theta}$.

We use pre-trained models to extract end-to-end features. ALBERT-base-v2 (Lan et al., 2019) is used for textual modality. We obtain the last hidden layer representation and project it to the Hilbert space with a linear layer: $\hat{r}_t = W_t \cdot ALBERT(T) + b_t$, where W_t and b_t are parameters. Then we normalize the outputs: $\vec{r}_t = \frac{\hat{r}_t}{||\hat{r}_t||^2}$. VGG (Simonyan and Zisserman, 2014) is used for visual and acoustic modalities. After the same processing as the textual modality, we obtain $\vec{r}_v$ and $\vec{r}_a$.

Pre-extracted features are obtained by handcrafted algorithms for visual (OpenFace2 (Baltrusaitis et al., 2018)) and acoustic (openSMILE (Eyben et al., 2010)) modalities. Motivated by previous work (Akhtar et al., 2019) that the sentiment polarity of words helps emotion recognition, we exploit a sentiment dictionary (Baccianella et al., 2010; Miller, 1995) to make use of sentiment polarity for the textual modality. Due to the advantage of capturing long-distance dependencies, the Transformer Encoder is used to encode these pre-extracted features.

Modal pure states $|\psi_t\rangle$, $|\psi_v\rangle$, $|\psi_a\rangle$ are constructed by formula (4) and the density matrices ρ_t, ρ_v, ρ_a are obtained by the outer product.

4.2 Adaptive Priority Learning

There are six fusion orders of three modalities. Based on the experimental results of previous work,

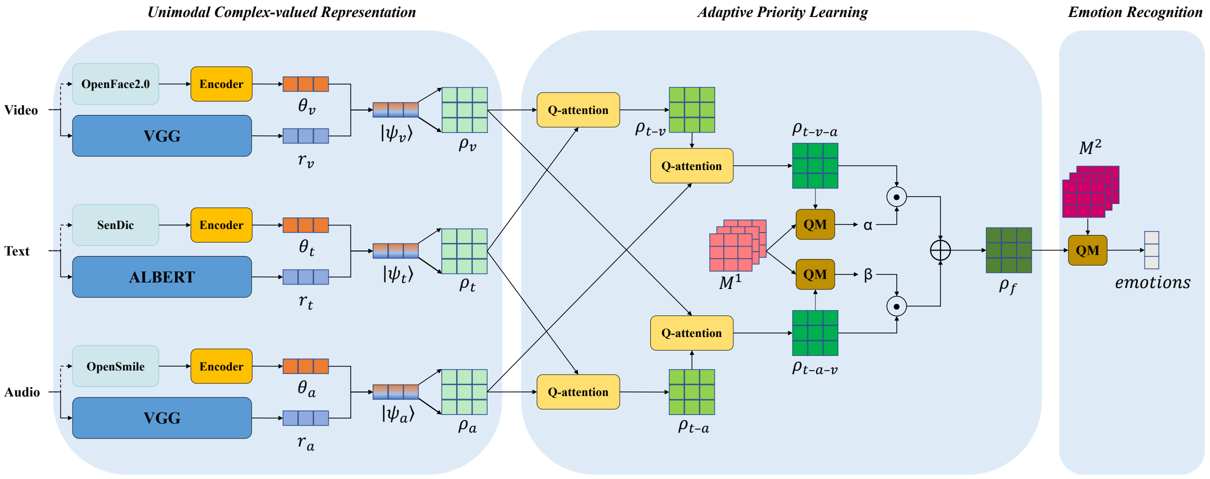


Figure 2: The overall architecture of QAP. ⊙ denotes the point-wise product and ⊕ denotes the element-wise addition. QM stands for Quantum Measurement. The dashed parts are not optimized through training.

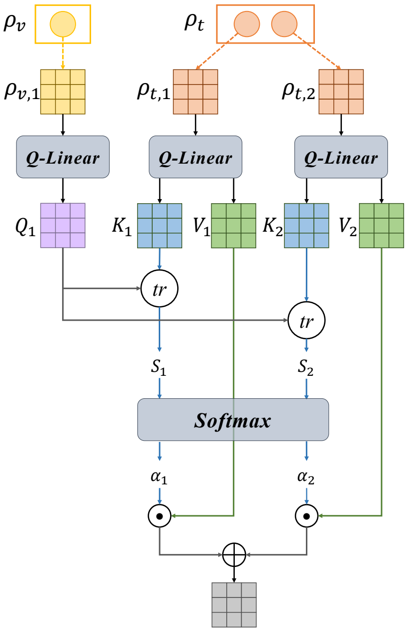


Figure 3: The main components of the Q-attention module.

textual modality usually contributes the most. Considering the computational cost, we only use two orders in our implementation: textual-visual-acoustic (t-v-a) and textual-acoustic-visual (t-a-v).

Taking the t-v-a order as an example, t and v are integrated first by Q-attention. The main process of Q-attention is shown in Figure 3. t is the basis, and v modality is to be added. ρ_t is fed into two Q-Linear layers to output K and V respectively, and ρ_v is also fed into a Q-Linear layer to output Q. Q-Linear is a linear layer designed for the density matrix analogous to quantum evolution:

$$K = U_1 \rho_t U_1^H, \tag{5}$$

$$V = U_2 \rho_t U_2^H, \tag{6}$$

$$Q = U_3 \rho_v U_3^H, \tag{7}$$

where U_1, U_2, U_3 are unitary matrices so K, V, Q are also density matrices. For pure states (vectors), attention scores can be calculated by the inner product, which cannot be directly applied to mixed states (density matrix). To solve this problem, we calculate the trace of the product of two density matrices:

$$\begin{aligned} tr(\rho_a \rho_b) &= tr(\sum_{i,j} p_i p_j |\psi_{a,i}\rangle\langle\psi_{a,i}|\psi_{b,j}\rangle\langle\psi_{b,j}|) \\ &= tr(\sum_{i,j} p_i p_j \langle\psi_{a,i}|\psi_{b,j}\rangle|\psi_{a,i}\rangle\langle\psi_{b,j}|) \\ &= \sum_{i,j} p_i p_j \langle\psi_{a,i}|\psi_{b,j}\rangle^2. \end{aligned} \tag{8}$$

Formula (8) proves that $tr(\rho_a, \rho_b)$ is the inner product weighted sum of the pure states. In fact, this is a generalization of the inner product from vectors to density matrices, called trace inner product (Balkır, 2014; Zhang et al., 2018). Therefore, we calculate the attention score between K and Q by trace inner product:

$$s_i = tr(K_i Q), \tag{9}$$

$$\alpha_i = Softmax(s_i). \tag{10}$$

Then, the output is obtained by weighted summation of V:

$$\hat{\rho}_{t\text{-}v} = \sum_i \alpha_i V_i, \tag{11}$$

where $\hat{\rho}_{\textit{t-v}}$ is the density matrix containing textual information and visual information. Inspired by Transformer (Vaswani et al., 2017), we also exploit the residual mechanism:

$$\hat{\hat{\rho}}_{\textit{t-v}} = \frac{1}{2}(\hat{\rho}_{\textit{t-v}} + Q), \tag{12}$$

$$\rho_{\textit{t-v}} = \frac{1}{2}(\hat{\hat{\rho}}_{\textit{t-v}} + Q \text{ - } Linear(\hat{\hat{\rho}}_{\textit{t-v}})), \tag{13}$$

where $\rho_{\textit{t-v}}$ is the fusion feature of textual and visual modalities. In addition, Q-attention is a multi-layer module. In the second and later layers, ρ_t is still the basis and used as K and V; while Q is the output of the previous layer and is continuously updated. So the whole process of Q-attention can be expressed by the following formula:

$$\rho_{\textit{t-v}} = Q\text{ - }attention(\rho_t, \rho_v). \tag{14}$$

Similar to the above procedure, acoustic modality can also be integrated by Q-attention. In the process, $\rho_{\textit{t-v}}$ is taken as K and V, and ρ_a as Q:

$$\rho_{\textit{t-v-a}} = Q\text{ - }attention(\rho_{\textit{t-v}}, \rho_a), \tag{15}$$

where $\rho_{\textit{t-v-a}}$ is the modal fusion feature in the order of t-v-a, and also a density matrix. In the same way, we can also obtain the modal fusion feature $\rho_{\textit{t-a-v}}$ in the order of t-a-v.

Then, a quantum measurement operator $M^1 = \{|m_j^1\rangle\}_{j=1}^n$ is utilized to select the most appropriate order for the current sample. The operator has n eigenstates so a n-dimensional probability distribution is calculated after the measurement of $\rho_{\textit{t-v-a}}$:

$$p_j^{\textit{t-v-a}} = tr(\rho_{\textit{t-v-a}} \, |m_j^1\rangle \, \langle m_j^1|). \tag{16}$$

We use a fully connected neural network to map the probability distribution to the weight of the t-v-a. $\rho_{\textit{t-a-v}}$ is also measured by M^1 and then the weight of the t-a-v order is obtained. We feed the two weights to a $Softmax$ layer and get α and β, where $\alpha + \beta = 1$. Finally, we sum the two density matrices:

$$\rho_f = \alpha \cdot \rho_{\textit{t-v-a}} + \beta \cdot \rho_{\textit{t-a-v}}, \tag{17}$$

where ρ_f is the multimodal fusion density matrix.

4.3 Emotion Recognition

We introduce another quantum measurement operator $M^2 = \{|m_j^2\rangle\}_{j=1}^n$ to recognize the emotions:

$$p_j^f = tr(\rho_f \, |m_j^2\rangle \, \langle m_j^2|), \tag{18}$$

$$p^e = FCN(p_f), \tag{19}$$

where $p^f = [p_1^f, p_2^f, \ldots, p_n^f]^T$ is an n-dimensional vector representing the probability distribution of each eigenstate and FCN is a fully connection neural network. $p^e = [p_1^e, p_2^e, \ldots, p_k^e]^T$ is the probability distribution of each emotion and k is the number of emotions.

During training, we use the BCEWithLogitsLoss function to calculate the loss.

5 Experiments

5.1 Datasets and Metrics

We conduct experiments to verify the performance of QAP on two widely used datasets: IEMOCAP and CMU-MOSEI. Both original datasets cannot be directly used for end-to-end training, so Dai et al. (2021) reconstructs these two datasets. After reconstruction, IEMOCAP contains 151 videos and 7,380 utterances. The content of each video is a dialogue between two professional actors according to the script. There are 6 emotion labels in IEMOCAP: {angry, happy, excited, sad, frustrated, neutral}. Each utterance only corresponds to one label. CMU-MOSEI is collected from the opinion videos on YouTube. The reorganized CMU-MOSEI contains 20,477 utterances and 6 emotion labels: {happy, sad, angry, fearful, disgusted, surprised}. Utterances in CMU-MOSEI may correspond to multiple labels. Following (Dai et al., 2020), we split the datasets, and the statistics of datasets are shown in Appendix A.

To comprehensively evaluate the performance of the method, we follow previous work (Dai et al., 2021) to use different evaluation indicators for the two datasets for fairness. For IEMOCAP, we calculate the accuracy and F1-score of each emotion and the average values. For CMU-MOSEI, we calculate the weighted accuracy and F1-score of each emotion and the average values.

5.2 Training Details

We use two optimizers during training. For unitary matrix parameters, we design an independent optimizer following Wisdom et al. (2016) to make these parameters always be unitary matrices in the training process. The optimization process is shown in Appendix B. For regular parameters, we use the Adam optimizer (Kingma and Ba, 2014). The experiments are run on a Tesla V100S GPU with 32GB of memory. There are about 58M parameters in our model. The time to run one epoch is less than one hour. We perform a grid search on

Models	Angry		Excited		Frustrated		Happy		Neutral		Sad		Average	
	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
LF-LSTM†	71.2	49.4	79.3	57.2	68.2	51.5	67.2	37.6	66.5	47.0	78.2	54.0	71.8	49.5
LF-TRANS†	81.9	50.7	85.3	57.3	60.5	49.3	85.2	37.6	72.4	49.7	87.4	57.4	78.8	50.3
EmoEmbs†	65.9	48.9	73.5	58.3	68.5	52.0	69.6	38.3	73.6	48.7	80.8	53.0	72.0	49.8
MulT†	77.9	60.7	76.9	58.0	72.4	57.0	80.0	46.8	74.9	53.7	83.5	65.4	77.6	56.9
AMOA	82.5	53.4	85.8	57.9	74.4	56.5	88.6	47.0	73.2	49.6	87.8	64.5	82.1	54.8
FE2E†	88.7	63.9	89.1	61.9	71.2	57.8	90.0	44.8	79.1	58.4	89.1	65.7	84.5	58.8
MESM†	88.2	62.8	88.3	61.2	74.9	58.4	89.5	47.3	77.0	52.0	88.6	62.2	84.4	57.4
QAP(ours)	**89.2**	**64.6**	**89.9**	**62.1**	**78.4**	**61.1**	**91.6**	**49.2**	**81.8**	**60.4**	**90.4**	**67.4**	**86.8**	**60.8**

Table 1: Results on the IEMOCAP dataset. † indicates that the results are from (Dai et al., 2021). Acc represents Accurary. Bolded numbers represent the best results.

Models	Angry		Disgusted		Fear		Happy		Sad		Surprised		Average	
	WAcc. ↑	F1 ↑	WAcc. ↑	F1 ↑	WAcc. ↑	F1 ↑	WAcc. ↑	F1 ↑	WAcc. ↑	F1 ↑	WAcc. ↑	F1 ↑	WAcc. ↑	F1 ↑
LF-LSTM†	64.5	47.1	70.5	49.8	61.7	22.2	61.3	73.2	63.4	47.2	57.1	20.6	63.1	43.3
LF-TRANS†	65.3	47.7	74.4	51.9	62.1	24.0	60.6	72.9	60.1	45.5	62.1	24.2	64.1	44.4
EmoEmbs†	66.8	49.4	69.6	48.7	63.8	23.4	61.2	71.9	60.5	47.5	63.3	24.0	64.2	44.2
MulT†	64.9	47.5	71.6	49.3	62.9	25.3	**67.2**	75.4	64.0	48.3	61.4	25.6	65.4	45.2
AMOA	66.4	47.5	74.9	52.2	62.0	25.1	62.6	73.4	63.8	47.2	64.3	26.5	65.7	45.3
FE2E†	67.0	49.6	77.7	57.1	63.8	26.8	65.4	72.6	65.2	49.0	66.7	29.1	67.6	47.4
MESM†	66.8	49.3	75.6	56.4	65.8	28.9	64.1	72.3	63.0	46.6	65.7	27.2	66.8	46.8
QAP(ours)	**68.7**	**52.4**	**78.8**	**59.6**	**67.3**	**30.3**	66.4	**75.9**	**65.4**	**50.1**	**66.7**	**31.3**	**68.9**	**49.9**

Table 2: Results on the CMU-MOSEI dataset. † indicates that the results are from Dai et al. (2021). WAcc represents Weighted Accuracy. Bolded numbers represent the best results.

the Valid set to select the hyper-parameters. The hyper-parameters are shown in Appendix C. For each experiment, we run three times and take the average.

5.3 Baselines

We compare QAP with several advanced multimodal emotion recognition models:

LF-LSTM: LSTM, a classical neural network, is used to encode modal features. It is a late fusion (LF) model.

LF-TRANS: The Transformer model is used to encode modal features and then the results are integrated. It is also a late fusion model.

EmoEmbs (Dai et al., 2020): This approach uses pre-trained word embeddings to represent emotion categories for textual data and transfer these embeddings into visual and acoustic spaces. EmoEmbs can directly adapt to unseen emotions in any modality and perform well in the zero-shot and few-shot scenarios.

MulT (Tsai et al., 2019): For modalities unaligned, MulT uses cross-modal attention to integrate modalities in pairs and does not pay attention to modal priority as above baselines.

AMOA (Li et al., 2022): Three modalities are integrated in a certain order and the global acoustic feature is introduced to enhance learning.

FE2E (Dai et al., 2021): FE2E is the first end-to-end model for MER, which uses pre-trained models to extract unimodal features and then fuses them.

MESM (Dai et al., 2021): Cross-modal attention and sparse CNN are utilized to integrate modalities and reduce computation based on FE2E.

5.4 Main Results

The experimental results on the IEMOCAP and CMU-MOSEI datasets are reported in Table 1 and Table 2, respectively. The results show that QAP outperforms all baseline models on average and most emotion categories. In general, QAP attains an improvement of **1%** - **3%** over other models, which indicates the advantage of QAP in MER.

The baseline models ignore the issue of inconsistent emotions, so they perform poorly in this situation. LF-LSTM, LS-TRANS, EmoEmbs and MulT are classic multimodal emotion recognition models but only use pre-extracted features. Besides, they treat modalities equally and do not pay attention to the important role of modal priority, so the performance is relatively poor. AMOA notices the importance of modal fusion order so the performance is improved compared with previous methods. However, AMOA cannot learn modal priority adaptively so the order is fixed. FE2E and MESM use end-to-end frameworks and can extract richer modal features, so they also perform well. But the two models also do not focus on modal priorities. QAP uses quantum states to model features so that modalities with inconsistent emotions can be effectively integrated. Besides, QAP learns modal priority adaptively and can adjust the modal fusion order based on priority, so outperforms all baselines.

5.5 Analysis

In order to further analyze the performance of QAP, we conduct extensive experiments on the IEMOCAP and CMU-MOSEI datasets.

Models	IEMOCAP		CMU-MOSEI	
	Acc.	F1	WAcc.	F1
QAP	**86.8**	**60.8**	**68.9**	**49.9**
- pure state	84.3	57.9	65.6	46.3
- concat	85.2	57.4	67.4	46.9
w/o phase	84.7	58.1	64.6	45.8
w/o SenDic	85.8	59.5	67.9	47.7

Table 3: Results of the ablation study of the complex-valued density matrix. - pure state means that the pure state is used to represent modal features instead of the density matrix. -concat means directly concatenating rather than using complex representation to combine two features. w/o phase means to remove pre-extracted features and only use real density matrix.

5.5.1 Effectiveness of complex-valued density matrix

To verify the role of the complex-valued density matrix, we change the unit of modal representation from the complex-valued density matrix to the pure state vector and conduct experiments. The results in Table 3 show that the performance of QAP decreases when the pure state is used. Besides, we try to directly concatenate end-to-end features and pre-extracted features rather than using complex representation. Experimental results show that this will also cause performance degradation.

We use complex value representation to combine pre-extracted features and end-to-end features. To verify the role of pre-extracted features, we remove the phase in the complex representation, that is, change the complex-valued matrix into the real-value matrix with only end-to-end features. As shown in Table 3, the addition of pre-extracted features makes a great contribution to the improvement of model performance. We introduce the sentiment dictionary into MER, and it is not used by other models, so we conduct an ablation study on SenDic individually. Results in the last row of Table 3 illustrate that the introduction of SenDic improves model performance.

Models	IEMOCAP		CMU-MOSEI	
	Acc.	F1	WAcc.	F1
QAP(Soft)	**86.8**	**60.8**	**68.9**	**49.9**
QAP(Hard)	85.3	58.7	66.9	47.5
-fixed(t-v-a)	83.4	57.1	66.5	45.7
-fixed(t-a-v)	83.2	56.9	64.8	44.7
-fixed(a-v-t)	81.8	56.2	63.9	44.0
-fixed(a-t-v)	82.4	57.2	64.2	43.5
-fixed(v-a-t)	80.5	55.8	63.1	43.0
-fixed(v-t-a)	80.8	56.1	63.9	43.4

Table 4: Experimental results of selection methods and fixed orders. Soft means to use $Soft\ selection$ and Hard means to use $Hard\ selection.$ -fixed(t-v-a) means that the fixed fusion order is t-v-a, and others are similar. The average results are reported.

Models	IEMOCAP		CMU-MOSEI	
	Acc.	F1	WAcc.	F1
QAP(t-a-v,t-v-a)	**86.8**	**60.8**	**68.9**	**49.9**
QAP(v-a-t,v-t-a)	85.1	58.4	66.7	46.3
QAP(a-t-v,a-v-t)	84.8	57.5	66.4	45.8
- 4 orders	86.5	61.3	67.4	49.0
- 6 orders	**87.3**	**61.6**	**69.7**	**50.6**

Table 5: Experimental results of orders reserved. The first three are cases where different two orders are reserved. - 4 orders means to use the four orders of t-v-a, t-a-v, v-a-t, and v-t-a. - 6 orders means to use all orders.

5.5.2 Effectiveness of adaptive-priority-learning fusion

Almost all baselines (except AMOA) do not integrate the three modalities in order, while QAP integrates the modalities in the order of modal priority, so the performance is better than all baselines, as shown in Table 1 and 2. In addition, compared with AMOA, our model adds a mechanism to adaptively adjust the fusion order and learn modal priority by a quantum measurement operator. To prove the effectiveness of this mechanism, we fix the modal fusion order in QAP and conduct experiments. The results are shown in Table 4, and we can see that no matter which fusion order is fixed, the model performance decreases. Therefore, any fusion order cannot be suitable for all samples and it is necessary to adaptively adjust the order according to different samples.

For the selection method of two fusion orders, we adopt the $Soft\ selection$ by default, which utilizes information in two fusion orders in a dynamic proportion. Besides, we also attempt to use $Hard\ selection$, that is, to discard the order with a lower score. The results in Table 4 show that QAP with $Soft\ selection$ performs better. The reason is that there is little difference between the contributions of acoustic and visual modalities in some samples, and both orders have positive contributions to emotion recognition.

In order to reduce the calculation, we only reserve two (t-v-a, t-a-v) of the six fusion orders based on the experimental results of previous work. We also utilize more orders and conduct experiments. The results in Table 5 show that the addition of more fusion orders does not significantly improve the performance with the increase of computation.

In addition, we also try to keep the other two orders and conduct experiments and the results are shown in Table 5. When we use the orders of t-a-v and t-v-a, QAP achieves the best performance, which indicates that our initial selection of the two fusion orders is appropriate.

Models	IEMOCAP		CMU-MOSEI	
	Acc.	F1	WAcc.	F1
QAP	**86.8**	**60.8**	**68.9**	**49.9**
QAP(non-orthogonal)	84.2	57.5	65.9	47.3
QAP(flatten)	84.8	58.1	65.7	46.9

Table 6: Experimental results of quantum measurement. non-orthogonal means that the eigenstates in measurement operator (M^1 or M^2) are not orthogonal to each other. flatten means to use FCN and softmax function instead of quantum measurement for order selection and emotion recognition.

5.5.3 Effectiveness of quantum measurement

In QAP, we use quantum measurement operators to collapse the density matrix ρ_f for classification (order selection and emotion recognition). This process unifies the entire classification process under a quantum-like framework and improves the interpretability of QAP. We also attempt to use two other non-quantum methods for classification to verify the effectiveness of quantum measurement. The first attempt is to use non-orthogonal eigenstates to form a measurement operator, which actually violates the concept of quantum measurement. The second attempt is to flatten rho_f to a one-dimensional vector, followed by a softmax function. The results in Table 6 show that the decreased performance of non-quantum methods reveals the superiority of quantum measurement.

Models	IEMOCAP		CMU-MOSEI	
	Acc.	F1	WAcc.	F1
QAP	**86.8**	**60.8**	**68.9**	**49.9**
w/o t	65.8	40.6	53.5	39.7
w/o v	80.3	54.9	62.2	42.6
w/o a	81.9	55.3	61.4	42.1

Table 7: Results of the ablation study of single modality. w/o means to remove this modality and only integrate the other two modalities.

5.5.4 Role of single modality

In MER, each modality plays an important role. And to verify the role, we separately remove one modality and conduct experiments. For example, when the textual modality is removed, the v-a and a-v orders are adopted and adaptively selected by a measurement operator. The results are shown in Table 7. When a modality is removed, the performance of QAP decreases in varying degrees. Specifically, when the textual modality is removed, the performance decreases most obviously, which is consistent with the results of previous work.

6 Conclusion

We propose QAP, a quantum-inspired adaptive-priority-learning model for multimodal emotion recognition. First, the quantum state is introduced to model the uncertainty of human emotion, which allows modalities with inconsistent emotions can be effectively integrated. Secondly, a novel mechanism Q-attention is designed to orderly integrate three modalities in a quantum-like framework. While selecting the appropriate fusion order, QAP learns modal priority adaptively. In this way, modalities make varying degrees of contributions based on priority. Experiments on two widely used datasets show that QAP establishes the new SOTA.

Limitations

We use the density matrix to represent modal features, and one of the advantages is that the matrix contains more information. However, the requirements for memory and large GPU resources also increase. Based on the best hyper-parameter setting, the shape of a pure state is 16×100×100, while the shape of a density matrix is 16×100×100×100. At the same time, the matrix will also increase the calculation and time cost. In future work, we will explore how to reduce the computational expense, and it is an idea to build the sparse density matrix.

References

Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multimodal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 370–379.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Esma Balkır. 2014. Using density matrices in a compositional distributional model of meaning. *Master's thesis, University of Oxford*.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Jerome R Busemeyer and Peter D Bruza. 2012. *Quantum models of cognition and decision*. Cambridge University Press.

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(4):335–359.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360.

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5305–5316.

Wenliang Dai, Zihan Liu, Tiezheng Yu, and Pascale Fung. 2020. Modality-transferable emotion embeddings for low-resource multimodal emotion recognition. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 269–280.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast opensource audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*.

Lauren Fell, Shahram Dehdashti, Peter Bruza, and Catarina Moreira. 2019. An experimental protocol to derive and validate a quantum model of decision-making. In *Annual Meeting of the Cognitive Science Society*.

Dimitrios Gkoumas, Qiuchi Li, Yijun Yu, and Dawei Song. 2021a. An entanglement-driven fusion neural network for video sentiment analysis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1736–1742. International Joint Conferences on Artificial Intelligence Organization.

Dimitris Gkoumas, Qiuchi Li, Shahram Dehdashti, Massimo Melucci, Yijun Yu, and Dawei Song. 2021b. Quantum cognitively motivated decision fusion for video sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 827–835.

Paul R Halmos. 2017. *Finite-dimensional vector spaces*. Courier Dover Publications.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7037–7041. IEEE.

Yongyu Jiang, Peng Zhang, Hui Gao, and Dawei Song. 2020. A quantum interference inspired neural matching model for ad-hoc retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–28.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Qiuchi Li, Dimitris Gkoumas, Alessandro Sordoni, Jian-Yun Nie, and Massimo Melucci. 2021. Quantum-inspired neural network for conversational emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13270–13278.

Qiuchi Li, Benyou Wang, and Massimo Melucci. 2019. Cnm: An interpretable complex-valued network for matching. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4139–4148.

Si Li and Yuexian Hou. 2021. Quantum-inspired model based on convolutional neural network for sentiment analysis. In *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 347–351. IEEE.

Ziming Li, Yan Zhou, Weibo Zhang, Yaxin Liu, Chuanpeng Yang, Zheng Lian, and Songlin Hu. 2022. Amoa: Global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7136–7146.

Yaochen Liu, Yazhou Zhang, Qiuchi Li, Benyou Wang, and Dawei Song. 2021. What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 871–880.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2019. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *AAAI Conference on Artificial Intelligence*.

Wasifur Rahman, M. Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2020:2359–2369.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Alessandro Sordoni, Jian-Yun Nie, and Yoshua Bengio. 2013. Modeling term dependencies with quantum language models for ir. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 653–662.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3722–3729. ACM.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

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*, pages 5998–6008.

Qinglan Wei, Xuling Huang, and Yuan Zhang. 2022. Fv2es: A fully end2end multimodal system for fast yet effective video emotion recognition inference. *IEEE Transactions on Broadcasting*.

Scott Wisdom, Thomas Powers, John Hershey, Jonathan Le Roux, and Les Atlas. 2016. Full-capacity unitary recurrent neural networks. *Advances in neural information processing systems*, 29.

Yang Wu, Zhenyu Zhang, Pai Peng, Yanyan Zhao, and Bing Qin. 2022. Leveraging multi-modal interactions among the intermediate representations of deep transformers for emotion recognition. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 101–109.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 1642–1651. ACM.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4400–4407. ACM.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Amir Zadeh and Paul Pu. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*.

Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1262–1274.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multimodal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3584–3593.

Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. 2018. End-to-end quantum-like language models with application to question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A The statistics of datasets.

	IEMOCAP					
	angry	happy	excited	sad	frustrated	neutral
Train	757	398	736	759	1298	1214
Valid	112	62	92	118	180	173
Test	234	135	213	207	371	321

	CMU-MOSEI					
	happy	sad	angry	fearful	disgusted	surprised
Train	7587	4026	3267	1263	2738	1465
Valid	945	509	318	169	273	197
Test	2220	1066	1015	371	744	393

Table 8: The statistics of IEMOCAP and CMU-MOSEI datasets.

B Optimization of unitary matrix

For a unitary matrix U used as a parameter, if its gradient is G, the optimization process is as follows:

$$A = G^H U - U^H G, \tag{20}$$

$$\hat{U} = (I + \frac{LR_u}{2}A)^{-1}(I - \frac{LR_u}{2}A)U, \tag{21}$$

where LR_u is the learning rate of the unitary matrix parameter optimizer. It can be proved that $\hat{U}$ is also a unitary matrix (Wisdom et al., 2016).

C Hyper-parameter Settings

	IEMOCAP	CMU-MOSEI
batch size	16	16
LR	3e-5	3e-5
LR_u	5e-5	8e-6
feature dim	100	100
sequence len	100	100

Table 9: Hyper-parameter settings of the two datasets. LR is the learning rate of general parameters and LR_u is the learning rate of unitary matrix parameters.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the section of 'Limitations'

☒ A2. Did you discuss any potential risks of your work?
Our work has not been found to have potential risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We propose a novel model in Section 4.

☑ B1. Did you cite the creators of artifacts you used?
Section 1 and 2 and 3 and 4 and 5.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The artifacts we use are free and public

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.

C ☑ Did you run computational experiments?

Section 5 and limitations.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 and Appendix C

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4 and 5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.