MMRBN: RULE-BASED NETWORK FOR MULTIMODAL EMOTION RECOGNITION

Xi Chen

School of Computer Science and Technology, Fudan University, Shanghai, China
chenxi19@fudan.edu.cn

ABSTRACT

Human emotion is usually expressed in multiple modalities, like audio and text. Multimodal methods can boost Emotion Recognition. However, the relationship between audio and text, and their roles in emotion expression have not been fully studied, and hence hinder Multimodal Emotion Recognition (MER). In this work, taking into consideration of the above two things, we propose two rules for MER, which are Rule 1: The audio module should be more expressive than the text module, and Rule 2: The single-modality emotional representation should be dynamically fused into the multimodal emotion representation. Following these two rules, we design the corresponding rule-based multimodal network (MMRBN). Experiment result on the public dataset demonstrates the effectiveness of our proposed rules and MMRBN.

Index Terms— Emotion Recognition, Multimodal Emotion Recognition, Attention Neural Network

1. INTRODUCTION

Emotion Recognition (ER) aims at detecting and recognizing human emotion, and it is essential for Human-Computer Interaction (HCI) to trigger accurate feedback. Since human emotion is often expressed in multiple modalities, like speech and spoken language, Multimodal Emotion Recognition (MER) has become a hot research point [1]. Like most previous work [2, 3], we focus on Audio-Text Multimodal Emotion Recognition (AT-MER).

Previous work builds up their model upon the emotion expression characteristics. They can be categorized into the following aspects. (a). Bridging in Contextual Information: Since the speaker's affective states are often influenced by the others in the dialogue, if historical contextual information available, effectively modeling and bridging in it can improve recognition performance [4, 2]. (b). Aligning text with audio: [5, 6] believe that the different time scale between audio representation and text representation hinders cross-modality information interaction. So they first align text with audio, and then process the multimodal input. (c). Utilizing powerful features: [7] and [3] utilize the pre-training model, which has good content understanding ability, as the basic embedding extractor.

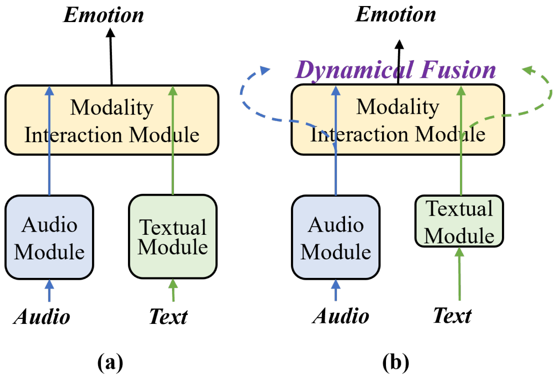

Fig. 1. This picture demonstrates the difference between models in the literature (a) and the model designed following our proposed rules (b).

Though numerous progress has been made by previous work, the audio modality and the text modality's relationship, and their roles in emotion expression have not been thoroughly explored. In this paper, we will study on these, propose two rules for designing MER models, and design the corresponding model.

One is on the **relationship between audio and text**. The text modality is not independent of the audio modality, the former is embedded in the latter naturally. Imagine that there is an oracle Speech Recognition (SER) system, the text information can be obtained from the audio recording directly. So, there are some modality-specific attributes in the audio, such as the tone, the speaking speed, and the loudness, while, the linguistic information provided by the text can also be explored from the audio. However, since using the text as an additional input can soften the audio module's burden on extracting the shared linguistic emotion attributes and enables us introducing knowledge from large language pre-training models, the AT-MER model is still necessary. In this way, in the AT-MER model, we proposed **Rule 1: The audio module should be more expressive than the text module.**

The other is on **how the audio and the text contribute to emotion expression**. The emotion expressed in each modality can be homogeneous or heterogeneous to the overall emotion the speaker expressed [8].

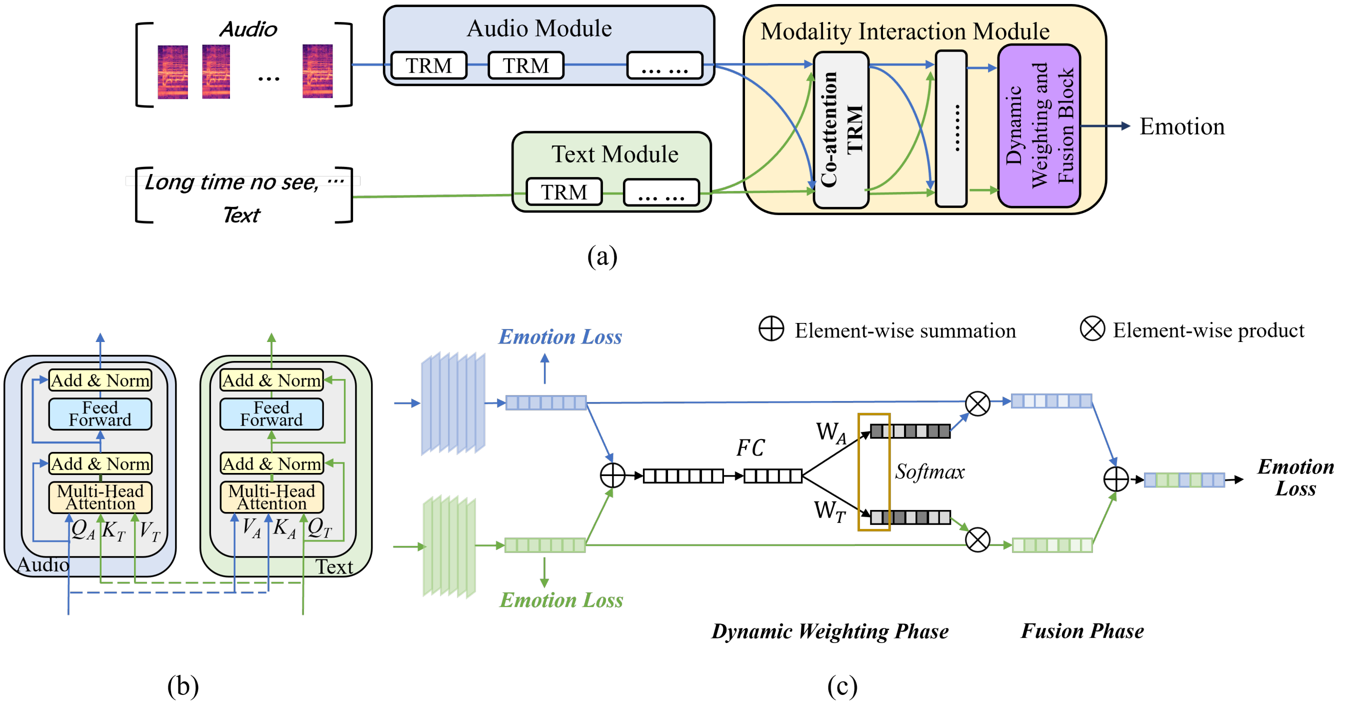

Fig. 2. (a). The overall architecture of MMRBN. (b). The Co-attention TRM. (c). The detail of Dynamic Weighting and Fusion Block.

When the emotion expressed in both modalities is homogeneous to the overall emotion, surplus information provided by both modalities can guide to better classification [9]. Example (a) presents an example of this kind, the word ”wonderful” in the text and the rising pitch and rising energy in the audio both contribute a lot to the overall joyful emotion.

When the emotion expressed in a certain single modality is heterogeneous to the overall emotion, the two modalities will contribute unequally to overall emotion expression, and the extent that each single-modality contributes is not firm. It depends case by case. Example(b) presents an example where the audio apparently contributes more, the text ”And I just want a million dollars” is neutral, while the rising pitch and volume in the audio illustrate the overall emotion is happy and excited.Example(c) presents an example where the text apparently contributes more, the tone is flat in the audio, while the words ”lucky” and ”wonderful” indicate that the speaker is happy.

Therefore, the extent each modality contributes to the overall emotion differs from case to case, and then we propose **Rule 2: Each single-modality emotion representation should be dynamically fused into the multimodal emotion representation**.

Overall, the difference between present models and those following our rules can be seen in Figure 1. And then we design our rule-based multimodal network (MMRBN). The experiment result shows that our MMRBN outperforms present state-of-the-art (SOTA) models.

2. METHODOLOGY

The overall architecture of our Rule-Based Network for Multimodal Emotion Recognition (MMRBN) is shown in Figure 2(a). It is composed of three modules, Audio Module (AM), Text Module (TM), and Modality Interaction Module (MI). AM and TM's design reflects the implementation of Rule 1, and it will be explained in Section 2.2. Dynamic Weighting and Fusion Block (DWFB) 's design, which is part of MI, reflects the implementation of Rule 2, and it will be explained in Section 2.3.2.

The input of the MMRBN is the audio embedding e_a and the text embedding e_t, output of the MMRBN is the emotion prediction $\hat{y}$.

2.1. Audio Module

The Audio Module (AM) is a stack of M Transformer's Encoder Layers (TRM) [10], because it enables global context modeling for sequential data and it works efficiently in a non-autoregressive mode. Each TRM is composed of a Multi-Head Attention sub-layer and a position-wise fully connected sub-layer, with the residual connection and layer normalization after each sub-layer. Input of AM is e_a, and output of AM is audio emotion representation r_a.

2.2. Text Module

Similar to the Audio Module, the Text Module is a stack of N TRM. Input of TM is e_t, and output of TM is text emotion representation r_t.

As Rule 1 states, The Audio Module should be more expressive than the Text Module. And we argue that the expressiveness of the model depends on its depth [11]. Therefore, for Rule 1's implementation, we strictly restrict MM-RBN with $M > N$.

2.3. Modality Interaction Module

The Modality Interaction Module (MI) is composed of two blocks. The former one is the Cross-attention Block (CAB) which enables representation space alignment and complementary attribute learning. The latter one is the Dynamic Selection and Fusion Block (DSFB), which deals with the multimodal representation fusion and the final decision-making, and it is our agent to implement Rule 2.

The input of MI are audio emotional representation e_a provided by AM and text emotional representation e_t provided by TM. Output is the emotion prediction $\hat{y}$.

2.3.1. Cross-attention Block

In addition to the implementation of Rule 2, for robust emotion recognition, the MI should deal with another two things. One is aligning representation space for e_a and e_t, because e_a and e_t are calculated separately without any knowledge interaction. The other is complementary attribute learning in both modalities.

The Cross-attention Block (CAB) is built up with a stack of L Co-attention Transformer Encoder Layers (Co-TRM) [12]. The Co-TRM, as shown in Figure 2(b), is actually composed of a pair of TRM, with one for each modality. These two TRMs swap Key(K) and Value(V) with each other. In this way, with the help of the Co-TRM's interior Multi-Head Attention Block, representation conditioned on the other modality can be obtained, which is actually the text-conditioned audio representation and the audio-conditioned text representation [12]. Therefore, representation space alignment and complementary attribute learning can be achieved.

As we can see, the input of CAB are e_a and e_t provided by AM and TM separately, the output of CAB are text-conditioned audio representation r_{A-t} and audio-conditioned text representation r_{T-a}.

2.3.2. Dynamic Weighting and Fusion Block

After the data being processed by CAB, we can have two streams of conditioned representation, r_{A-t} and r_{T-a}. How to fuse them and make the final decision is stated in Rule 2. Hence, our Dynamic Weighting and Fusion Block (DWFB) is designed to implement it.

The DWFB, shown in Figure 2(c), works in two phases. They are the Dynamic Weighting Phase, which dynamically weights each conditioned single-modality representation's contribution to the final multimodal representation, and the Fusion Phase, which conducts the final modality fusion based on the weights provided by the Dynamic Weighting Phase.

Dynamic Weighting Phase: To achieve dynamic weighting, referencing to [13], the basic idea is to use attention to control the information flow from r_{A-t} and r_{T-a} to the final multimodal emotional representation.

First, the Global Average Pooling (GAP) is used to get the corresponding utterance-level conditioned single-modality representation $R_A \in \mathbb{R}^d$ and $R_T \in \mathbb{R}^d$, where d is the representation dimension,

Second, all the attributes in both modalities should be integrated together before calculating dynamic weights for each modality. Therefore, the compact uni-representation $G_U \in \mathbb{R}^{d'}$ is calculated as:

$$G_U = ReLU(FC(R_A \oplus R_T)). \tag{1}$$

In Equation 1, the element-wise addition $\oplus$ enables information integration; viewing the representation channels as attributes, the fully connected layer $FC(\cdot)$ and the ReLU activate function $ReLU(\cdot)$ enable attribute interaction. And, a ration r is used to control the attribute interaction rate, by restricting $r \cdot d' = d$.

Third, two projection matrixes, $W_A \in \mathbb{R}^{d \times d'}$ and $W_T \in \mathbb{R}^{d \times d'}$ are employed to calculate the importance of each attribute in each modality.

$$I_A = W_A \cdot G_U, I_T \quad = W_T \cdot G_U, \tag{2}$$

where $I_A \in \mathbb{R}^d$ and $I_A \in \mathbb{R}^d$.

Finally, the dynamic attribute-wise cross-attention weight $Attn_A$ and $Attn_T$ are calculated using the element-wise softmax function between I_A and I_T.

Fusion Phase: The final multimodal emotion representation can be calculated by re-weighting and adding R_A and R_T together.

$$R_M = R_A \otimes Attn_A \oplus R_T \otimes Attn_T, \tag{3}$$

where $\otimes$ is the element-wise manipulation.

2.4. Loss

The final loss function L is composed of three parts, with L_A and L_T, added upon R_A and R_T, and L_M, added upon R_M.

$$\begin{aligned} L_T = \alpha \cdot L_A + \beta \cdot L_T + \gamma \cdot L_M \\ s.t. \alpha + \beta + \gamma = 1, \end{aligned} \tag{4}$$

where $\alpha, \beta, and, \gamma$ are loss weights.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Metric

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [14] is used for fair evaluation. There are nine categorical emotion annotations. Like most of the previous work, we only use samples from four categories: anger, sadness, happiness (samples labeled excitement are merged with happiness), and neutral. Overall, the dataset we used contains 5,531 samples, with 1,103, 1,636, 1,084, and 1,708 samples from each of the above-mentioned categories.

We conduct 5-fold cross-validation. The evaluation metrics we used are Weighted Accuracy (WA) and Unweighted Accuracy (UA). WA is the overall accuracy and UA is the average of the class-level accuracy. When there is a class imbalance problem, WA puts more weight on the class with more data, while UA weighs each class equally.

3.2. Implementation

As for preprocessing, the embedding of the audio and the text are extracted from Hubert [15] and Bert [16] respectively. what's more, before sending the embedding into our MMRBN, a trainable linear layer is added to resize the embedding into the same dimension.

As for the model, the depth M of AM is set to be 4, the depth N of TM is set to be 2, which indeed satisfies our Rule 1.

3.3. Ablation Study

In this subsection, we would like to validate the reasonability of the proposed two rules, and the effectiveness of our MMRBN. We conduct experiments on seven models, they are (1). AM_6, which is a single-modality audio model and is composed of 6 TRM; (2). TM_6, which is a single-modality text model and is composed of 6 TRM; (3). MMRBN_2 (w/o Rule 1), which is almost the same as our MMRBN without taking into consideration of Rule 1, where the depth of AM and TM are both 2; (4). MMRBN_4 (w/o Rule 1), which is almost the same as our MMRBN without taking into consideration of Rule 1, where the depth of AM and TM are both 4; (5). MMRBN_conca (w/o Rule 2), which is almost the same as our MMRBN without taking consideration of Rule 2, where the DWFB is discarded and the concatenation operation is used to fuse the two single-modality representations; (6). MMRBN_add (w/o Rule 2), which is almost the same as our MMRBN without taking consideration of Rule 2, where the DWFB is discarded and the element-wise addition operation is used to fuse the two single-modality representation; (7). MMRBN, which is the proposed.

Table 1. Ablation study result.

ID	Model	WA	UA
(1)	AM_6	0.678	0.702
(2)	TM_6	0.684	0.709
(3)	MMRBN_2 (w/o Rule 1)	0.735	0.755
(4)	MMRBN_4 (w/o Rule 1)	0.739	0.759
(5)	MMRBN_conca (w/o Rule 2)	0.738	0.746
(6)	MMRBN_add (w/o Rule 2)	0.733	0.749
(7)	MMRBN	0.760	0.766

The ablation study result can be seen in Table 1. Experiment results between (1), (2) and (3)-(7) can demonstrate the effectiveness of using multimodal methods in Emotion recognition. Experiment results between (3), (4) and (7) can demonstrate the effectiveness of Rule 1. Experiment results between (5), (6) and (7) can demonstrate the effectiveness of Rule 2. These all agree with our hypothesis. Besides, experiment results between (3), (4) and (5), (6) can demonstrate Rule 2 is a little more effective than Rule 1.

3.4. Result

Table 2. Comparison result with SOTA methods, where A-T means both audio and text modality are used as input, A means only audio is used as input.

Year	Modality	Model	WA	UA
2022	A	[17]	0.698	0.711
2021	A-T	[18]	0.717	0.75
2022	A-T	[19]	0.743	0.754
2023	A-T	[20]	0.741	0.754
2023	A-T	[21]	0.752	0.764
2023	A-T	MMRBN	0.760	0.766

As shown in Table 2, we compare our MMRBN with SOTA methods, and our model achieves the best performance. Among all, [17] performs badly slightly, because it is a single-modality model. Among all the other multimodal methods, [18] performs worst, because the modality interaction method used is too simple, which is merely an attention operation.

4. CONCLUSION

In this work, we propose two rules for multimodal emotion recognition, and we design the corresponding MMRBN. The experiment result shows that MMRBN achieves SOTA performance. In the future, we will focus more on multimodal representation fusion and how to effectively utilize more modality information.

5. REFERENCES

[1] Zihan Zhao, Yu Wang, and Yanfeng Wang, "Knowledge-aware bayesian co-attention for multimodal emotion recognition," in *ICASSP*. IEEE, 2023, pp. 1–5.

[2] Wen Wu, Chao Zhang, and Philip C Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *ICASSP*. IEEE, 2021, pp. 6269–6273.

[3] Zihan Zhao, Yanfeng Wang, and Yu Wang, "Multi-level fusion of wav2vec 2.0 and bert for multimodal emotion recognition," *arXiv preprint arXiv:2207.04697*, 2022.

[4] Darshana Priyasad, Tharindu Fernando, Sridha Sridharan, Simon Denman, and Clinton Fookes, "Dual memory fusion for multimodal speech emotion recognition," *INTERSPEECH*, pp. 4543–4547, 2023.

[5] 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 *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2018, vol. 2018, p. 2225.

[6] Guang Shen, Riwei Lai, Rui Chen, Yu Zhang, Kejia Zhang, Qilong Han, and Hongtao Song, "Wise: Word-level interaction-based multimodal fusion for speech emotion recognition.," in *Interspeech*, 2020, pp. 369–373.

[7] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP*. IEEE, 2022, pp. 6922–6926.

[8] Feiyu Chen, Zhengxiao Sun, Deqiang Ouyang, Xueliang Liu, and Jie Shao, "Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1064–1073.

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

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[11] Wonjae Kim, Bokyung Son, and Ildoo Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

[12] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[13] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[14] 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," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[15] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, vol. 29, pp. 3451–3460, 2021.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP*. IEEE, 2022, pp. 7367–7371.

[18] Puneet Kumar, Vishesh Kaushik, and Balasubramanian Raman, "Towards the explainability of multimodal speech emotion recognition.," in *Interspeech*, 2021, pp. 1748–1752.

[19] Weidong Chen, Xiaofeng Xing, Xiangmin Xu, Jichen Yang, and Jianxin Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP*. IEEE, 2022, pp. 6897–6901.

[20] Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li, "Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities," in *ICASSP*. IEEE, 2023, pp. 1–5.

[21] Suzhen Wang, Yifeng Ma, and Yu Ding, "Exploring complementary features in multi-modal speech emotion recognition," in *ICASSP*. IEEE, 2023, pp. 1–5.