ISSN 2616-5775 Vol. 4, Issue 8: 10-18, DOI: 10.25236/AJCIS.2021.040803

# GME-Dialogue-NET: Gated Multi-modal Sentiment Analysis Model Based on Fusion Mechanism

**Meng Yang, Yegang Li, Hao Zhang**

*School of Computer Science and Technology, Shandong University of Technology, Zibo 255049, Shandong, China*

***Abstract:*** *In comparison with the single mode, the utilization of multi-mode information of text, video and audio could lead to more accurate sentiment analysis. GME-Dialogue-NET, a gated multi-modal sentiment analysis model, is raised for the multi-modal emotion prediction and sentiment analysis. The model judges whether the audio or video modal is the noise through GME (Gated Multi-modal Embedding, GME) and then accepts or refuses the modal information based on the judgement. The model uses the Attention Mechanism of context vector to allocate more attention to the context with greater relevance to the current sentence. GME-Dialogue-NET divides participants of the dialogue into speaker and listener to better capture the dependence between emotion and state. It raises that the fusion mechanism CPA (Circulant-Pairwise Attention, CPA) could pay effective attention with different degrees on different modals to attain more helpful emotional and sentimental representation and thus make emotion prediction and sentiment analysis. Compared with the current model, both the weighted accuracy and the F1 score of emotion prediction were improved, especially for the three emotions of sadness, anger and excitement. In the sentiment regression task, the comparison between GME-Dialogue-NET with current advanced model Multilogue-Net shows that MAE (Mean absolute error, MAE) of GME-Dialogue-NET reduces by 0.1 percentage and the Pearson Correlation Coefficient (R) of GME-Dialogue-NET rises by 0.11 percentage.*

***Keywords:*** *Natural language processing, Multi-modal sentiment analysis, Multi-modal fusion mechanism*

## 1. Introduction

With the growth of various social platforms in quantity, more and more ways are exposed to people to express their emotions on the Internet[1][2][4][5][8][15]. The study of human emotions has also evolved from single-mode to multi-mode. Multimodal emotion analysis refers to the use of computers and related technologies to obtain information from multimodal data such as language, sound and image to analyze the emotions expressed by people, which is one of the more active research fields in natural language processing and has received extensive attention and research.

In 2008, Datcu and Rothkrantz[3] proposed a dual-mode semantic data fusion model, which combined visual and auditory information to identify six prototype emotions, among which the detection accuracy of surprise emotion category reached 88.67%. In 2010, Wollmer et al[6] proposed a multi-modal emotion detection and emotion analysis technology based on feature-level fusion. In this paper, we propose for the first time that bidirectional long and short-term memory (BLSTM) networks can be used to model the evolution of emotions in conversations, taking long-distance information into account. BLSTM network method is superior to traditional classification techniques (such as hidden Markov model or support vector machine). Poria et al[7] proposed a model based on LSTM in 2017. Compared with the technique proposed by Wollmer, the overall framework of the model is elaborated in more detail and several variations are proposed. The interdependencies between utterances can be used to capture contextual information. Zadeh et al proposed TFN[8] and MFN[9] in 2017 and 2018 respectively, and Ghosal et al.[10]in 2018 proposed the paired attention mechanism, which are studies on the fusion mechanism in multimodal emotion analysis. In 2018, Majumder et al[11] put forward the Dialogue RNN model, which uses the gated loop unit (GRU)[12] to effectively track the state and current context of the participants and distinguish the participants in the Dialogue so as to capture contextual information more effectively. However, Dialogue RNN fails to use an effective fusion mechanism (the fusion method is to connect the extracted feature representations of each mode) and pays insufficient attention to the correlation between multiple modes, which hinders its performance. In

ISSN 2616-5775 Vol. 4, Issue 8: 10-18, DOI: 10.25236/AJCIS.2021.040803

order to solve the problems in Dialogue RNN, Shenoy et al.[13]proposed the model multilogue-NET in 2020. Multilogue-net uses the pin-attention mechanism to fuse the modal information.

Generally speaking, the main challenges for multimodal sentiment analysis include whether the context can be captured effectively, whether the information of each mode is redundant or noisy, and whether each mode can be effectively fused. In light of the problems, innovation points are proposed:

1) A Gated multi-modal Embedding (GME) is proposed for receiving or rejecting audio or video modal information.

2) A new fusion method CPA (Circulant-Pairwise Attention) was proposed.Based on this fusion method, a multi-modal sentiment analysis model, GME-Dialogue-NET, is proposed to produce advanced performance.

## 2. Introduction to the model

The proposed model, GME-Dialogue-NET, has two modules: Gated multimodal embedding and dialogue-NET modules respectively.

Gated multimodal embedding: Gated multimodal embedding is implemented through a deep network.Gated multimodal embedding uses a multimodal input gate to determine how much video or audio information for each word needs to be retained or discarded.

The dialogue-net module: Conducting sentiment prediction and sentiment analysis for every sentence of each participant.This module distinguishes the participants in the Dialogue, which can more effectively capture the context in the Dialogue and track the dependence of emotions on the state. Gated Recurrent Units (GRU) can be adopted in Dialogue-net to obtain such information.

## 3. GME (Gated Multi-modal Embedding, GME)

The module introduces an on/off input gate controller to accept or reject audio or video messages. Each video clip is divided into T time steps, and each time step corresponds to a word. The multi-modal representation of the word T is: $x_t^t, x_t^a, x_t^v$ .

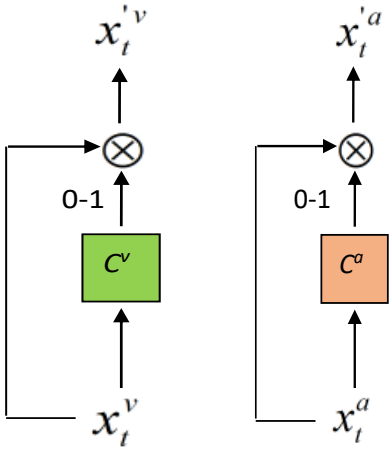



*Figure 1: The overall framework of Gated Multi-modal Embedding*

A controller $C^a$ can be used with a weight of $\theta_a$ to determine the audio mode on/off. The controller $C^v$ can be used with the weight of $\theta_v$ , the switch used to determine the video mode. The controller uses deep neural network $C^a(\bullet;\theta_a)$ and $C^v(\bullet;\theta_v)$ for implementation. Input $x_t^a, x_t^v$ for output as a probability value $c_t^a, c_t^v$ between 0 and 1 as a label. The output of the controller mimics the behavior of rejecting or accepting modal information, with 0 indicating total rejection and 1 indicating total acceptance. (See Figure 1). The formula is as follows:

$$x_t'^a = c_t^a \bullet x_t^a = C^a\left(x_t^a;\theta_a\right)\bullet x_t^a \tag{1}$$

ISSN 2616-5775 Vol. 4, Issue 8: 10-18, DOI: 10.25236/AJCIS.2021.040803

$$x_t^{'v} = c_t^v \bullet x_t^v = C^v\left(x_t^v;\theta_v\right)\bullet x_t^v \tag{2}$$

After calculation, the multi-modal expression of each word is $x_t^t, x_t^{'a}, x_t^{'v}$. The mean value of the feature representation of all words in each statement was calculated to obtain the statement-level feature representation $t_t \in R^{D_t}, a_t \in R^{D_a}, v_t \in R^{D_v}$, which was used as the input of the module Dialogue-net to make emotion analysis for each statement.

## 4. The Dialogue-Net module

### *4.1. Problem description*

Supposing that the participant in the conversation is p, which can be noted as $p_1$, $p_2$, …$p_P$. Each word spoken by each participant is recorded as $u_1$, $u_2$, …, $u_N$. In light of the timestamp t and the utterances $u_t$ of the participants, each available mode (text T, audio A and video V) has an independent feature representation, which is respectively $t_t \in R^{D_t}, a_t \in R^{D_a}, v_t \in R^{D_v}$.

### *4.2. Overall Framework*

Assuming that the emotion of each sentence depends on the follows: 1) The speaker's current words or the listener's current expression; 2. Current context; 3. Current status of participants; 4. Participants' previous emotions. Figure 2 is the overall framework of the dialogue-NET module, m∈{t, a, v}.

### *4.3. Context GRU (GRUc)*

Each mode has one GRUc, and m modes have m GRUCs in quantity. The GRUc of a particular mode co-encodes the characteristic representation of the modal input statement and the state representation of the participant, producing a valid context representation. A timestamp $t_{t-1}$ statement indicates that $t_{t-1}$, $a_{t-1}$, or $v_{t-1}$ changes the participants' state from $s_{t-1}^m\left(s_{t-1}^t, s_{t-1}^a, s_{t-1}^v\right)$ to state $s_t^m\left(s_t^t, s_t^a, s_t^v\right)$.GRUc takes this change and outputs a fixed-size vector.

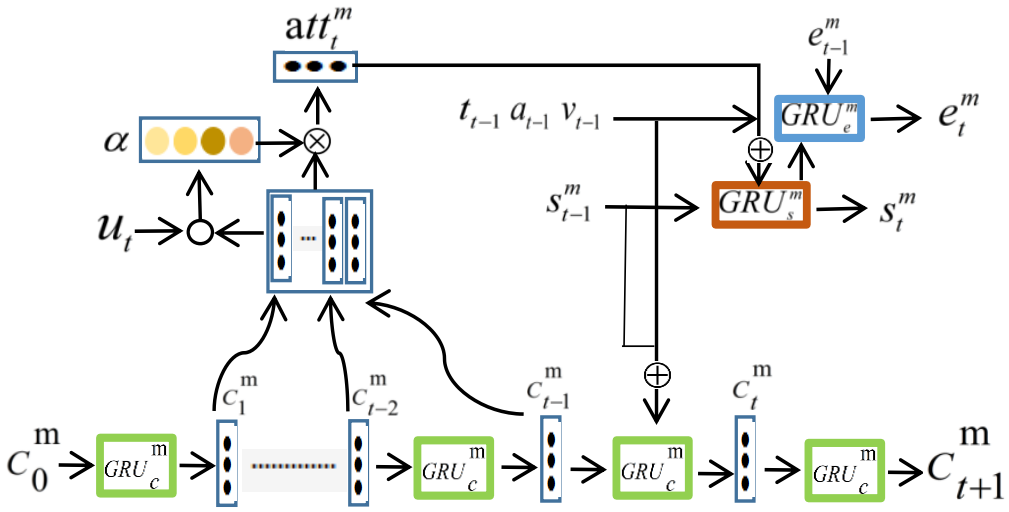



*Figure 2: The overall framework of Dialogue-Net*

$$c_t^t = GRU_c\left(c_{t-1}^t, \left(t_{t-1} \oplus s_{t-1}^t\right)\right) \tag{3}$$

$$c_t^a = GRU_c\left(c_{t-1}^a, \left(a_{t-1} \oplus s_{t-1}^a\right)\right) \tag{4}$$

$$c_t^v = GRU_c\left(c_{t-1}^v, \left(v_{t-1} \oplus s_{t-1}^v\right)\right) \tag{5}$$

Among them, $D_{t,a,v}$ are the sizes of the text, audio, and video features of the statement respectively.

ISSN 2616-5775 Vol. 4, Issue 8: 10-18, DOI: 10.25236/AJCIS.2021.040803

$D_s \in R^{D_s}$ refers to the magnitude of the state vector $s_{t-1}^t$, $s_{t-1}^a$ and $s_{t-1}^v$. $D_c \in R^{D_c}$ denotes the magnitude of the context vector $c_t^t$, $c_t^a$ and $c_t^v$. $\oplus$ indicates a connection operation.

### *4.4. Status GRU (GRUs)*

Each mode has a state GRU for each participant. The state vectors of all participants in the conversation are initialized to empty vectors. Figure 3 (a) is the update of speaker state $GRU_{s,l}$; Figure 3(b) Update of listener state GRUs, s, m∈{t, a, v}.

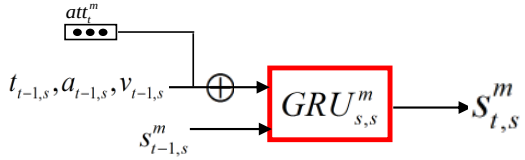



*Figure 3(a): The update of the speaker state GRU (GRUs, s)*

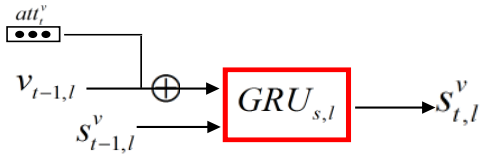



*Figure 3(b): The update of the listener state GRU (GRUs, l)*

#### *4.4.1. Speaker status GRU (GRUs, s)*

Speakers respond to context, which is tracked by the context GRUc.The context at timestamp *t* is simply a note of all the context vectors from timestamp 1 to timestamp *t*-1. m∈{t, a, v} is described by the following formula:

$$\alpha = \text{softmax}\left(u_t^T W_\alpha \left[c_1^m, c_2^m, ..., c_{t-1}^m\right]\right) \tag{6}$$

$$att_t^m = \alpha\left[c_1^m, c_2^m, ..., c_{t-1}^m\right]^T \tag{7}$$

Among them, $u_t^T \in \left\{t_t^T, a_t^T, v_t^T\right\}$, $W_\alpha \in R^{D_m \times D_c}$, $att_t^m \in R^{D_c}$. $c_1^m, c_2^m, ..., c_{t-1}^m$ in formula 6 is the context representation of all statements up to a certain time stamp. Attention score is calculated for all context representation, and higher attention score is assigned to the statements with a high degree of emotional relevance. Formula 7 uses α to amplify related statements to obtain context $att_t^m$.

At the timestamp t, the model $GRU_{s,s}^m$ uses the characteristic representation of the statement at time *t*-1 and the context to update the speaker state from $s_{t-1,s}^{t,a,v}$ to $s_{t,s}^{t,a,v}$.

$$s_{t,s}^t = GRU_{s,s}\left(s_{t-1,s}^t, \left(t_{t-1,s} \oplus att_t^t\right)\right) \tag{8}$$

$$s_{t,s}^a = GRU_{s,s}\left(s_{t-1,s}^a, \left(a_{t-1,s} \oplus att_t^a\right)\right) \tag{9}$$

$$s_{t,s}^v = GRU_{s,s}\left(s_{t-1,s}^v, \left(v_{t-1,s} \oplus att_t^v\right)\right) \tag{10}$$

Among them, $D_s \in R^{D_s}$ refers to the size of the state vector $s_{t,s}^t$, $s_{t,s}^a$ and $s_{t,s}^v$.

#### *4.4.2. Listener status GRU (GRUs, L)*

The listener's state changes due to the speaker's words, which is mainly reflected in facial expression, facial muscles, facial movements, etc.[14]. These information comes from the video mode. In timestamp t, the video feature representation and context are used by $GRU_{s,l}$ to update the listener status representation.

$$s_{t,l}^v = GRU_{s,l}\left(s_{t-1,l}^v, \left(v_{t-1,l} \oplus att_t^v\right)\right) \tag{11}$$

ISSN 2616-5775 Vol. 4, Issue 8: 10-18, DOI: 10.25236/AJCIS.2021.040803

Among them, $v_{t-1,l} \in R^{D_V}$ and $D_s$ demonstrate the size of the listener's state vector $s_{t,l}^v$.

Two state GRUs encode the state information of all participants in the current conversation.

### *4.5. Emotional GRU helped (GRUe)*

An emotional GRU (GRUe) is actually a decoder for state encoding that outputs the emotional or emotional representation of a particular statement. The updated formula of emotion and emotion expression is as follows:

$$e_t^m = GRU_e\left(e_{t-1}^m, s_{t,s}^m\right) \tag{12}$$

Among them, multimodal m∈{t, a, v}. $D_e$ is the size of all emotion (emotion) vectors.

### *4.6. Circulant-Pairwise Attention, CPA)*

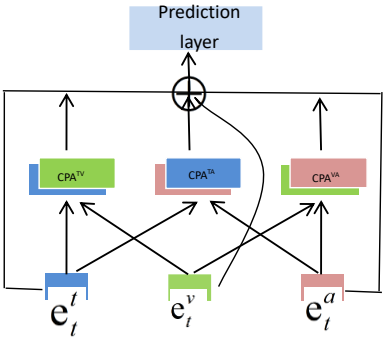



*Figure 4: Circulant-Pairwise Attention used as the fusion mechanism*

Each mode produces one emotion or emotion representation for each timestamped statement, and there are m emotion or emotion representations for m available modes in quantity. The CPA fusion mechanism is used to fuse m kinds of emotion or emotion representation of each sentence. The CPA fusion mechanism algorithm is made to transform the emotion and emotion representation of each mode into a cyclic matrix, and then carry out paired attention. This model has three available modes, so CPA calculates three pairs, which are respectively ($e^t$, $e^v$), ($e^t$, $e^a$), ($e^a$, $e^v$). which is depicted in Figure 4.

Taking text mode $e^t$ and video mode $e^v$ as examples for calculation:

$$M_1 = circ\left(e^t\right), M_2 = circ\left(e^v\right) \tag{13}$$

$$B_1 = \frac{1}{n}\sum_{i=1}^{n} m_{1i} \bullet (e^v)^T, B_2 = \frac{1}{n}\sum_{i=1}^{n} m_{2i} \bullet (e^t)^T \tag{14}$$

$$N_1 = soft\max\left(B_1\right), N_2 = soft\max\left(B_2\right) \tag{15}$$

$$O_1 = N_1 \bullet e^t, O_2 = N_2 \bullet e^v \tag{16}$$

$$A_1 = O_1 \odot e^v, A_2 = O_2 \odot e^t \tag{17}$$

$$pairwise\left(e^v, e^t\right) = A_1 \oplus A_2 \tag{18}$$

Among them, $circ(e^t)$ and $circ(e^v)$ in Formula 13 and 14 are used to transformed $e^t$ and $e^v$ i nto a cyclic matrix, that is, each row of the vector is shifted by one element to obtain the ma trix. The Pairwise algorithm was thoroughly analyzed in the article [13] Contextual Inter-modal Attention for Multi-Modal Sentiment Analysis. ⊙ stands for multiplying by elements.

$$pw = pw\left(e^v, e^t\right) \oplus pw\left(e^a, e^t\right) \oplus pw\left(e^a, e^v\right) \tag{19}$$

$$L_t = pw \oplus e_t^t \oplus e_t^a \oplus e_t^v \tag{20}$$

ISSN 2616-5775 Vol. 4, Issue 8: 10-18, DOI: 10.25236/AJCIS.2021.040803

Among them, $pw(e^v,e^t)$ represents $pairwise(e^v,e^t)$.

### *4.7. Emotion prediction or emotion analysis*

The fusion features obtained by formula (20) in Section 3.6 represent that $L_t$ is input to a full connected layer, followed by a *tanh* layer or *soft*max layer, depending on whether it is emotion prediction or emotion analysis.

Emotion analysis: The affective forecasting value when the time stamp t is output. The formula of the output layer is as follows:

$$P_{sentiment(t)} = \tanh\left(W_L L_t\right) \tag{21}$$

Among them, $W_L \in R^{9D_e \times 1}$.

Emotion prediction: $L_t$ can be input to a fully connected layer is followed by a *soft*max layer to calculate the probability of 6 emotions.

$$l_t = \mathrm{Re}LU\left(W_l L_t + b_l\right) \tag{22}$$

$$P_t = soft\max\left(W_{s\max} l_t + b_{s\max}\right) \tag{23}$$

$$P_{emotion(t)} = \arg\max\left(P_t[i]\right) \tag{24}$$

Among them, $W_l \in R^{D_l \times 9D_e}$; $b_l \in R^{D_l}$; $W_{s\max} \in R^{c \times D_l}$; $b_{s\max} \in R^c$; $P_t \in R^c$.

## 5. Variants of the model GME-Dialogue-NET

There are three variants of the model GME-Dialogue-NET: (1) Dialogue-NET: To remove the first module of the model and to retain the second module dialogue-net. (2) GME-Dialogue-NET (NA): Removing the simple attention of context vector and replacing it with direct connection of all context vector. (3) GME-Dialogue-NET (NF), the variant removes the fusion mechanism of the model.

## 6. Experiments

### *6.1. Data set*

The model was tested on multimodal datasets CMU-MOSI and IEMOCAP.

CMU-MOSI dataset[16]: Film comments based videos are collected, and each video is divided into multiple segments. Each clip has an emotional label value y, which is evaluated on a continuous range between -3 and + 3. In addition, the data set was divided into two categories, positive and negative emotions.

IEMOCAP dataset [17]: This dataset contains the interactions of 10 actors and actresses in an emotion-binary dialogue. There are 6 categories of emotion labels in the data set.

The data set is divided into training set, validation set and test set, with each part accounting for 75%, 15% and 10%, respectively.

### *6.2. Feature Extraction*

The feature representation of three modes is extracted, namely text feature representation, audio feature representation and video feature representation. The extraction of text feature representation is achieved with the use of pre-trained word embedding (Glox.840b.300d)[18] to transform scripts in data set videos into word vectors. Audio feature representation was extracted using COVAREP [19,20]. For video, Facet and OpenFace [21,22,23] are used to extract a set of features.

### *6.3. Experimental results*

In order to evaluate the model, Weighted Accuracy and F1 score were used to evaluate the emotion

ISSN 2616-5775 Vol. 4, Issue 8: 10-18, DOI: 10.25236/AJCIS.2021.040803

prediction task. The sex of weighted precision ratio accuracy to each class.

You can be more sensitive. The ACCURACY and recall of the classification model are considered by using F1 score, which makes the classification result more meaningful.

The affective analysis task is conducted with the adoption of mean absolute error (MAE) and Pearson correlation coefficient (R) for evaluation, which can get the difference between the predicted value and the true value, so as to better evaluate the model.

#### *6.3.1. Comparison with current advanced models*

Table 1 and Table 2 show the model GME-Dialogue-NET and its variants compared to current advanced models.

Table 1 demonstrates a comparison of GME-Dialogue-NET and its variants with other advanced models on the data set CMU-MOSI. In terms of dichotomous accuracy, GME-Dialogue-NET is higher than multilogue-NET at 0.16 percentage points, and the scores of F1 is 0.21 percentage points higher. In the Pearson correlation coefficient of regression index, GME-Dialogue-NET is higher than DialogueRNN at 0.12 percentage points, and the mean absolute error is 0.12 percentage points lower.

Table 2 shows the results of sentiment classification on data set IEMOCAP and the comparison of four models including CNN, DialogueRNN, Multilogue-NET, and GME-Dialogue-NET. For the detection of happy emotion, multilogue-NET is 0.1 percentage points higher than GME-Dialogue-NET, and the model GME-Dialogue-NET has the highest performance in other emotion detection. In the measurement of anger, GME-Dialogue-NET outperforms DialogueRNN by 5.2 percentage points on the weighted accuracy index. In addition, the detection of excitement and depression also has obvious performance improvement.

*Table 1: GME-Dialogue-NET performance on CMU-MOSI compared with other models*

| Approachs | CMU-MOSI | | | |
|---|---|---|---|---|
| | A2 | F1 | MAE | r |
| CNN | 74.89 | 75.03 | 0.82 | 0.38 |
| MMMU-BA | 80.03 | 79.84 | 0.67 | 0.43 |
| DialogueRNN | 80.57 | 79.78 | 0.69 | 0.47 |
| Multilogue-Net | 80.87 | 79.81 | 0.67 | 0.48 |
| Dialogue-Net | 80.97 | 79.88 | 0.63 | 0.52 |
| GME-Dialogue-Net (NF) | 80.94 | 79.87 | 0.62 | 0.53 |
| GME-Dialogue-Net | **81.03** | **80.02** | **0.57** | **0.59** |

*Table 2: GME-Dialogue-NET performance on IEMOCAP compared with previous models*

| Approachs | IEMOCAP | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Happy | | Sad | | Neutral | | Angry | | Excited | | Frustrated | |
| | WA | F1 | WA | F1 | WA | F1 | WA | F1 | WA | F1 | WA | F1 |
| CNN | 62.6 | 72.8 | 69.1 | 76.6 | 62.1 | 89.9 | 66.3 | 66.3 | 60.4 | 66.9 | 53.7 | 85.5 |
| DialogueRNN | 82.2 | 80.9 | 90.3 | 87.4 | 89.7 | 87.1 | 70.1 | 68.5 | 76.1 | 74.6 | 87.5 | 84.1 |
| Multilogue-Net | **83.3** | 81.7 | 92.6 | 87.3 | 89.8 | 88.3 | 73.4 | 70.9 | 77.3 | 76.7 | 88.5 | 84.7 |
| GME-Dialogue-Net | 83.2 | **81.9** | **92.7** | **89.6** | **89.8** | **89.1** | **75.3** | **72.5** | **79.3** | **78.7** | **89.4** | **86.3** |

#### *6.3.2. Gated multimodal embedding analysis*

Gated multimodal embedding contributes to multimodal fusion, and variant dialist-net is susceptible to noise modal information, which is also confirmed in Table 1. Figure 5, from the data set CMU-MOSI, shows a speaker covering his mouth while speaking the emotion-expressing word "cute." And the variant dialogue-net cannot reject the video modal information that makes no sense at this moment and make the wrong emotional analysis. GME-Dialogue-NET rejects the video modal information for the word "cute" and gives a more realistic affective forecasting. This shows that the model GME-Dialogue-NET can make the right decision based on whether the current video mode or audio mode matches the text mode information.

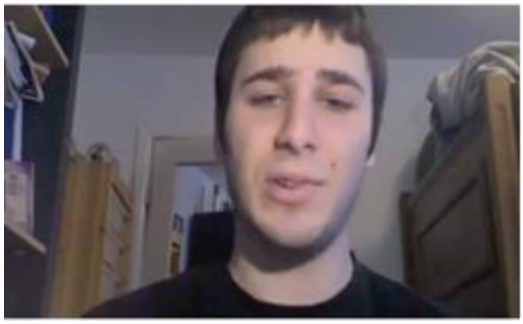
ISSN 2616-5775 Vol. 4, Issue 8: 10-18, DOI: 10.25236/AJCIS.2021.040803

*Dialogue-net affective prediction: -0.69*
*GME-Dialogue-NET affective prediction: 1.87*
*True emotion label: 3.0*

*Figure 5: Successful example for gated multi-modal embedding*

#### *6.3.3. Analysis of fusion mechanism*

Model GME-Dialogue-NET is made with more attention to modal information with greater emotional value through the fusion mechanism CPA. Figure 6 is an example of a fusion mechanism that works. It comes from the data set Cmu-MOSI, where the text is the only actor who can really sell their lines is Erin, and the emotional information provided by the text is vague and ambiguous. The video mode provides information that the speaker looks sad when he or she says the sentence. The model GME-Dialogue-NET learns that in this example the video mode is more valuable for predicting real emotion, thus making the correct emotion analysis. Thus, when all modes provide consistent, powerful emotional information, the model GME-Dialogue-NET makes the right judgment. When there is ambiguity among various modal information, GME-Dialogue-Net can use the fusion mechanism CPF to pay more attention to the modal information that is more valuable to real emotion and make correct prediction.



*The video mode: Looks sad*
*Mae-dialogue-ne (NF) affective prediction: 1.86*
*GME-Dialogue-NET affective prediction: -0.3*
*True emotion label: -1.0*

*Figure 6: Successful example for fusion mechanism*

## 7. Conclusion

GME-Dialogue-Net is a gated multimodal sentiment analysis model based on fusion mechanism. Experiments show that the simple attention mechanism of gated multi-modal embedding context vector and the application of fusion mechanism can improve the performance of the model.

## Acknowledgments

This paper is supported by the National Natural Science Foundation of China (No. 61671064).

ISSN 2616-5775 Vol. 4, Issue 8: 10-18, DOI: 10.25236/AJCIS.2021.040803

**References**

*[1] Richards J M, Butler E A, Gross J J. Emotion regulation in romantic relationships: The cognitive consequences of concealing feelings[J]. Journal of social and personal relationships, 2003, 20 (5): 599-620*

*[2] He Jun, Liu Yue, He Zhongwen.Research process of multimodal emotion recognition[J]. Application Research of Computers. 2018, 35 (11): 3201-3205.*

*[3] Datcu, D., Rothkrantz, L.Semantic audio-visual data fusion for automatic emotion recognition. [J] Euromedia'2008.*

*[4] Kanade T, Cohn J F, Tian Y. Comprehensive database for facial expression analysis[C]//Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580). IEEE, 2000: 46-53*

*[5] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]//Ninth european conference on speech communication and technology. 2005.*

*[6] Wöllmer M, Metallinou A, Eyben F, et al. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling[C]//Proc. INTERSPEECH 2010.The Prefecture of Grenoble, in France: ISCA Press, 2010: 2362-2365.*

*[7] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th annual meeting of the association for computational linguistics. Stroudsburg, PA: ACL Press, 2017: 873-883..*

*[8] Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J]. ArXiv, 2017, 1707.07250*

*[9] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence.Menlo Park, CA: AAAI Press, 2018: 5634–5641.*

*[10] Ghosal D, Akhtar S M, Chauhan D, Poria S, et al. Contextual inter-modal attention for multi-modal sentiment analysis[C]//In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA: ACL Press.2018: 3454-3466.*

*[11] Majumder N, Poria S, Hazarika D, et al. Dialoguernn: An attentive rnn for emotion detection in conversations[C]//Proceedings of the AAAI Conference on Artificial Intelligence.Menlo Park, CA: AAAI Press, 2019: 6818-6825.*

*[12] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. ArXiv, 2014, 1406.1078.*

*[13] Shenoy A, Sardana A. Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation[J]. ArXiv preprint ArXiv: 2020, 2002.08267,.*

*[14] Ekman P. Facial expression and emotion[J].American psychologist, 1993, 48 (4): 384-392.*

*[15] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. ArXiv, 2014, 1412.6980.*

*[16] Zadeh A, Zellers R, Pincus E,, et al.MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos[J]. arXiv preprint arXiv: 2016, 1606.06259.*

*[17] Busso C, Bulut M, Lee C C, et al.IEMOCAP: Interactive emotional dyadic motion capture database[J].Language resources and evaluative, 2008, 42 (4): 335–359*

*[18] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.*

*[19] Degottex G, Kane J, Drugman T, et al. COVAREP—A collaborative voice analysis repository for speech technologies[C]// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, 960–964*

*[20] Tao Huawei, Cha Cheng, Liang Ruiyu, et al. Spectrogram feature extraction algorithm for speech emotion recognition[J].JOURNAL OF SOUTHEAST UNIVERSITY (Natural Science Edition), 2015, 45 (05): 817-821.*

*[21] Baltrušaitis T, Robinson P, Morency L P. Openface: an open source facial behavior analysis toolkit[C]//2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016: 1-10.*

*[22] Zadeh A, Chong Lim Y, Baltrusaitis T, et al. Convolutional experts constrained local model for 3d facial landmark detection[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. IEEE 2017: 2519-2528.*

*[23] Zhu Q, Yeh M C, Cheng K T, et al. Fast human detection using a cascade of histograms of oriented gradients[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 2: 1491.*