Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-11234-y

1182: DEEP PROCESSING OF MULTIMEDIA DATA

Multimodal sentiment analysis with asymmetric window multi-attentions

Helang Lai[1,2] · **Xueming Yan**[3,4]

Received: 18 June 2020 / Revised: 18 June 2021 / Accepted: 8 July 2021 /
Published online: 23 July 2021

Abstract
Multimodal sentiment analysis is an actively developing field of research. The main research problem in this domain is to model both intra-modality and inter-modality dynamics. However, most of the current work cannot do well with these two aspects of dynamics. In this study, we introduce a novel model to achieve this. The novelty of our model is to represent the asymmetric weights of contexts at a particular timestamp using asymmetric windows. Further, multiple separate attentions are performed on the contexts, producing an updated representation of the particular timestamp. Each representation corresponding to one of the modes multiplies a weight vector controlled by a neural network. All multiplied results are merged with an addition operation. Experiments on the MOSI dataset show our model outperforms the compared methods.

Keywords Sentiment analysis · Asymmetric window · Multi-attentions · Neural network · Multimodal

1 Introduction

Multimodal research as an emerging research field of artificial intelligence (AI) has exhibited great advantages in a variety of tasks, such as speech recognition [15, 43], emotion recognition [16, 19], media description [30], and sentiment analysis [1, 17, 23, 49]. Multimodal sentiment analysis focuses on generalizing text-based sentiment analysis to

Helang Lai and Xueming Yan had contributed equally to this work.

✉ Xueming Yan
xueming126@126.com

[1] Guangdong Justice Police Vocational College, Guangzhou 510520, China

[2] School of Computer Science, South China Normal University, Guangzhou 510631, China

[3] Guangzhou Key laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou 510000, China

[4] School of Information Science and Technology, School of Cyber Security, Guangdong University of Foreign Studies, Guangzhou 510000, China

opinionated videos. It involves learning and analyzing rich representations from data across multiple modalities [2]. The multimodal data is collected from diverse perspectives and has heterogeneous properties. Each modality in multimodal sentiment analysis can have its own particular representation space and contain some information that other modes do not refer to [46]. There is a requisite for computational approaches that can integrally model multi-modal data. It can provide robust predictions and obtain outstanding generalization ability through exploring the consistency and complementary properties of different modalities to integrate the multi-modal data [33].

Although the analysis of this multimodal sentiment is natural for humans, it is difficult for AI to achieve comprehensive and accurate integration to classify the sentiments as humans do. The central challenge is to model both the intra-modality and inter-modality dynamics. The intra-modality dynamics are those interactions that come from the same modality and are independent of other modalities. The inter-modality dynamics are those interactions that span across both the different modalities and time [47]. Figure 1 illustrates two short examples extracted from the MOSI dataset, which respectively involve the intra-modality and inter-modality dynamics.

To capture these dual dynamics in multimodal sentiment analysis, most of the current research either performs feature-level fusion or decision-level fusion [45]. Feature-level

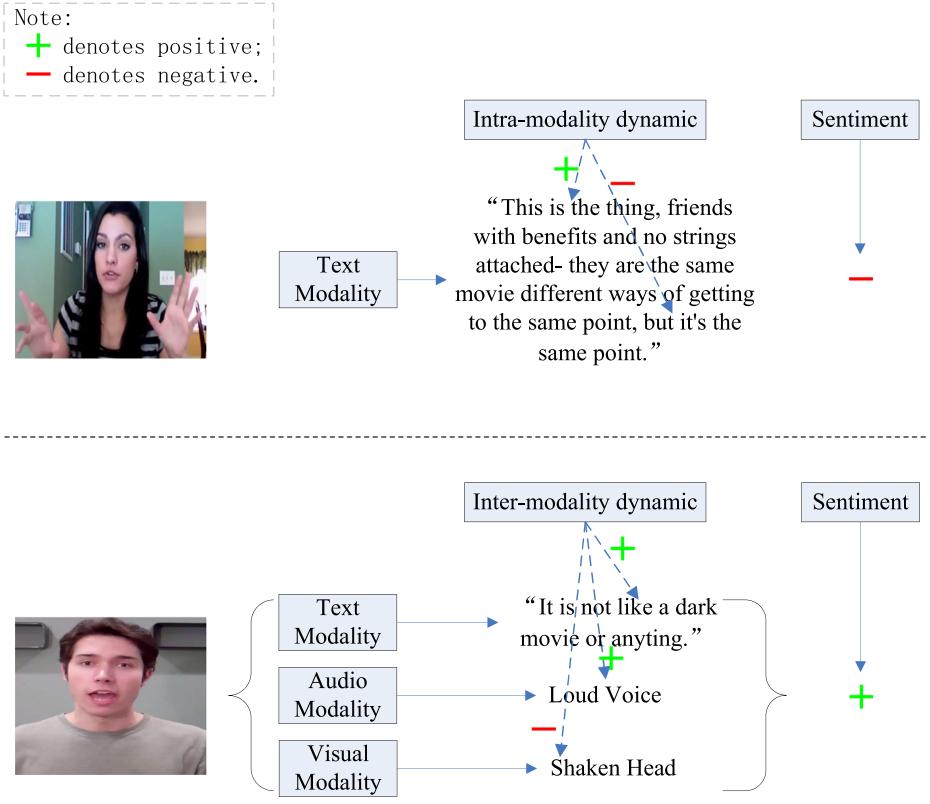

Fig. 1 Two examples from the MOSI dataset, have the intra-modality and inter-modality dynamics, respectively

fusion approaches are simply to concatenate multimodal features at the input level [17, 21, 25]. Decision-level fusion approaches firstly train unimodal classifiers independently and finally perform decision voting [36, 48]. Besides, Fukui et al. [8] employed Compact Bilinear Pooling (CBP) over the outer product of visual and linguistic representations to capture the interactions between vision and language for visual question answering. Zadeh et al. introduced Tensors Fusion Network (TFN) to compute a multimodal tensor representation [45], used the Multi-attention Recurrent Network (MARN) mechanism to model both intra-modality and inter-modality dynamics [47], and learned these dynamics through using Memory Fusion Network (MFN) [46]. However, none of these studies focuses on the asymmetric weights of the historical and future contexts at a particular timestamp of the input data when performing an attention mechanism.

In this paper, we propose a novel method called asymmetric window multi-attentions (AWMA) neural network, to model both the intra-modality and inter-modality dynamics in multimodal sentiment analysis. Our AWMA is based on the assumption: the historical and future contexts at a particular timestamp of the input data under different modes should have different weights. Further, these weights can be represented with sliding asymmetric windows when being performed an attention mechanism. AWMA has three main components. At the first component, text, audio, and visual modality, and the concatenation of all modalities, a total of four aspects as inputs, are encoded respectively using four separate gated recurrent unit (GRU) modules, to model the intra-modality dynamics. Then, the outputs of all the GRU modules are sent to the second component. This component is the core of our model, which consists of four asymmetric window attention (AWA) modules. Each AWA module uses sliding asymmetric windows to perform an attention mechanism on the contexts of a particular timestamp. This is to realize our hypothesis that these asymmetric windows can distinguish and represent the weights of the contexts under different modes. Furthermore, all the outputs of the four AWA modules are conveyed to the third layer. This component is called the inter-modality attention (IMA) module, which is to model the inter-modality dynamics. This module is for the aim that each modality and the concatenation of all modalities, a total of four aspects, should have different weight vectors to represent their importance. Each weight vector is controlled by a neural network and does a dot product with the output of its assigned AWA module. Then, all the dot product results are merged using an addition operation. To this, both intra-modality and inter-modality dynamics are integrated by our model. Finally, the output of the third layer is sent to the fully connected (FC) layer for the dimensional reduction of the prediction.

The main contributions of our paper are as follows:

- We propose an AWMA model to distinguish the asymmetric weights of the historical and future contexts at a particular timestamp of the input data. To our knowledge, it is the first time to represent these implicit asymmetric weights with asymmetric windows.
- We evaluate our approach with experiments on the public multimodal dataset MOSI. The experimental results demonstrate our AWMA model obtains competitive performance in comparison with the baseline models.
- We perform a brief ablation study of four AWA modules of our method.

The remainder of the paper is organized as follows: Section 2 briefly discusses previous work; Section 3 describes our model in detail; Section 4 provides experimental methodology; Section 5 presents the result and discussion; finally, Section 6 concludes the paper.

2 Related work

Multimodal sentiment analysis as a branch of affective computing research [22] is rapidly attracting the attention of both within academia, because of the many new challenges, and in the business world, due to the remarkable benefits to be had from financial and political forecasting [7, 38], community detection [3], human communication comprehension [47], and dialogue systems [14, 42], etc. Research in this field has facilitated us to utilize complementary information present in multimodal data, so that we have discovered the dependency of information on multiple modalities. However, modeling both the intra-modality and inter-modality dynamics in sentiment analysis is a crucial challenge.

Most of the previous approaches focused on feature-level fusion and decision-level fusion. Feature-level fusion is also called early fusion, a technique that combines the features extracted from different modalities into a 'joint vector' before any classification operations are performed [11, 24, 29]. Several works [12, 25, 26, 36] use this method to concatenate the features as input to a learning model. Decision level fusion also called late fusion, is to model and classify each modality separately. All the unimodal results are integrated at the end of the process by choosing suitable metrics, such as majority voting or weighted averaging [11, 24]. For example, the works [5, 9, 18, 37, 50] all use this fusion strategy. Although some degree of success for modeling the fusion problems of multimodal data is achieved, these two categories of methods do not directly account for both intra-modality and inter-modality dynamics and, have several drawbacks. In feature-level fusion, intra-modality dynamics are potentially suppressed [41], which results in losing out on the context and temporal dependencies within each modality. In decision-level fusion, inter-modality dynamics are not modeled effectively because independent models are built for each intra-modality, making inter-modality dynamics intricate to capture with simple weighted averaging or other similar decision methods.

Based on the above shortcomings, several recent works focus on multi-view learning approaches that model both intra-modality and inter-modality dynamics. Wang et al. [35] introduced several deep multi-view representation learning models, but they did not explore the applicability of sequence learning problems. Xu et al. [39] and Jia et al. [40] proposed to employ conventional LSTM as extensions of multi-view representation learning. Vinyals et al. [34] also used LSTM in the decoder module to obtain image sentence representations. Ren et al. [28] designed a multimodal LSTM to leverage all view representations. Song et al. [31, 32] proposed multi-view learning from multiple different modalities as the extensions of Hidden Markov Models and Hidden Conditional Random Fields. Rajagopalan et al. [27] also developed a multi-view LSTM model where the LSTM memory was partitioned into different components for different modalities. Zadeh et al. [46] learned intra-modality interactions in isolation through assigning an LSTM function to each modality, while identified inter-modality interactions using a special attention mechanism called the Delta-memory Attention Network (DMAN) and summarized them through time with a Multi-view Gated Memory. Besides, Zadeh et al. [47] discovered interactions between modalities through time using a neural component called the Multi-attention Block (MAB) and storing them in the hybrid memory of a recurrent component called the Long-short Term Hybrid Memory (LSTHM).

While these recent approaches can learn both intra-modality and inter-modality dynamics to some extent, they do not explicitly differentiate the asymmetric weights of the historical and future contexts at a particular timestamp of the input data. Our work is a follow-up to previous research [47]. Although our proposed model also utilizes multiple attention

mechanisms, our method is different as we employ four AWA modules to perform asymmetric window multi-attentions.

3 Asymmetric window multi-attentions neural network

The asymmetric window multi-attentions (AWMA) neural network consists of three main components: gated recurrent unit (GRU) module, asymmetric window attention (AWA) module, and inter-modality attention (IMA) module.

The GRU module and AWA module are employed to capture the intra-modality dynamics in multimodal sentiment analysis, and the IMA module is to model the inter-modality dynamics. We combine the three modules step by step. First, the input sequence is passed through the GRU module, then the AWA module, and finally the IMA module. The novelty of our proposed method is to use asymmetric windows to represent the asymmetric weights of historical and future contexts at a particular timestamp of the input data.

Figure 2 illustrates the overall model. In the following, we describe these components and their inputs in detail.

3.1 GRU module

The input to AWMA is a multi-modality sequence. In this paper, sequences can consist of text (T), audio (A), visual (V), and their concatenation ($T \oplus A \oplus V$), a total of four aspects. The m^{th} aspect input data is of the form $X^m = \{x_1^m, x_2^m, ..., x_i^m, ..., x_{N-1}^m, x_N^m\}$, where $x_i^m \in R^{d_{X^m}}$, x_i^m is the input at time i and $R^{d_{X^m}}$ is the dimensionality of the m^{th} aspect input data. $m \in M$. $M = \{T, A, V, T \oplus A \oplus V\}$.

For the computations of different intra-modality dynamics of the input data under M different aspects, we employ four independent GRU modules that are based on gated recurrent units (GRUs). Similar to an LSTM [10] in recurrent neural network (RNN), the GRU

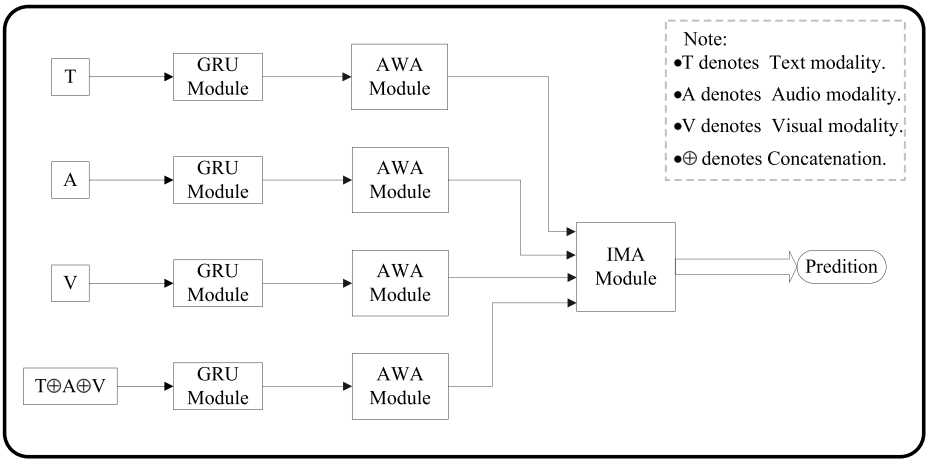

Fig. 2 Overall architecture of AWMA

introduced by Cho et al. [4] is a simpler gating mechanism with similar computation performance, which has a memory to store a representation of its input through time. Note that these four GRU modules allow different sequences to have different input, memory, and output shapes. Each aspect sequential information is as the input of its assigned GRU module.

At any time step t of a temporal sequence, the GRU controls the combination criteria with the t^{th} input u_t and previous state s_{t-1} by computing two gates, r_t (reset gate) and z_t (update gate). The computations are:

$$z_t = \sigma(V^z \cdot u_t + W^z \cdot s_{t-1} + b^z) \tag{1}$$

$$r_t = \sigma(V^r \cdot u_t + W^r \cdot s_{t-1} + b^r) \tag{2}$$

$$h_t = tanh(V^h \cdot u_t + W^h \cdot (s_{t-1} \otimes r_t) + b^h) \tag{3}$$

$$s_t = (1 - z_t) \otimes h_t + z_t \otimes s_{t-1} \tag{4}$$

Here, V, W and b are trainable parameters and $\otimes$ represents element-wise multiplication.

3.2 AWA module

The AWA module uses asymmetric windows to represent the asymmetric weights of the historical and future contexts at a particular timestamp of the input data. This input data to the AWA module is the output of the GRU module. The goal of the AWA module is to highlight the most important part of the interaction process. Figure 3 presents an overview of this module.

Let us assume the data X to be a sequence corresponding to one aspect. X_t denotes the information at the timestamp t. The historical contexts of X_t are information occurring before X_t. The future contexts of X_t are information occurring later X_t. In this paper, we constrain the historical and future contexts with asymmetric windows. For example, we use $W_{his} = 3$ to denote the sliding window of historical contexts and $W_{fut} = 2$ to denote the sliding window of future contexts. We concatenate both the historical and future contexts and mark the concatenation as C_{hf}. As shown in Fig. 3, $C_{hf} = \{X_{t-3}, X_{t-2}, X_{t-1}, X_{t+1}, X_{t+2}\}$. As the attention mechanism can focus on the most

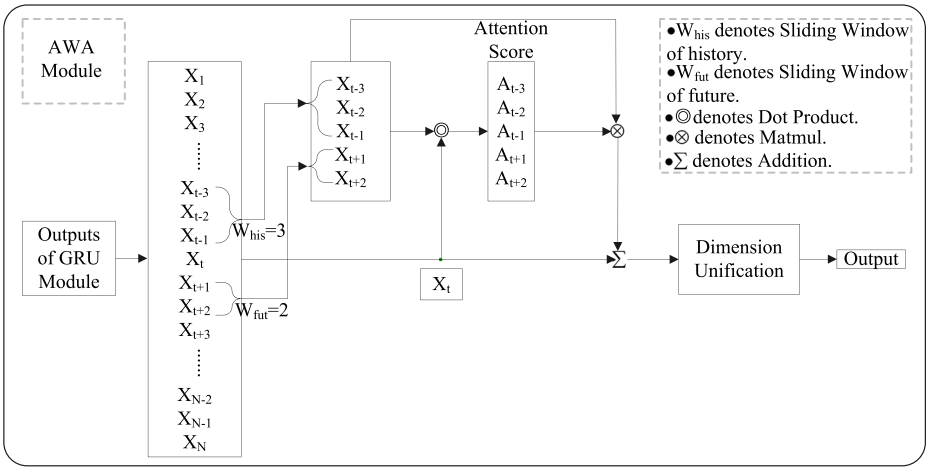

Fig. 3 Overview of AWA module

important parts relevant to the classification, we perform multi-separate attentions on the contexts under different aspects. One module performs a separate attention mechanism corresponding to one aspect, resulting in an attention vector A of dimension $R^{W_{his}+W_{fut}}$ for C_{hf}. The computations are:

$$A = softmax((C_{hf})^T \cdot X_t) \tag{5}$$

Where, $softmax(x_j) = e^{x_j} / \sum_k e^{x_k}$, and the attention vector A is a probability distribution over C_{hf}. The j^{th} normalized score of vector A represents the relevance of j^{th} contextual cell with respect to X_t. The A and C_{hf} are then used to find a weighted contextual representation. We do this as follows:

$$attC_{hf} = C_{hf} \cdot A \tag{6}$$

To this, the representation $attC_{hf}$ contains a weighted contextual summary accumulated from the historical and future contexts. The representation of X_t can be updated by adding the weighted contextual representation as:

$$X_t = X_t + attC_{hf} \tag{7}$$

Finally, this updated X_t is sent to the IMA module after the operation of dimension unification.

3.3 IMA module

Figure 4 demonstrates an overview of the IMA module. All the outputs of AWA modules are conveyed to this module to model the inter-modality dynamics. Let us assume the inputs of this module to be T_{att}, A_{att}, V_{att}, and $(TAV)_{att}$. The concatenation of all inputs is marked as $concatAll = T_{att} \oplus A_{att} \oplus V_{att} \oplus (TAV)_{att}$. For the computations of different weight vectors under different aspects, we employ four neural networks as controllers. Each neural network has the same input data $concatAll$, but it works independently during the training phase. We mark the four neural networks as NN_1, NN_2, NN_3, and NN_4, respectively, and

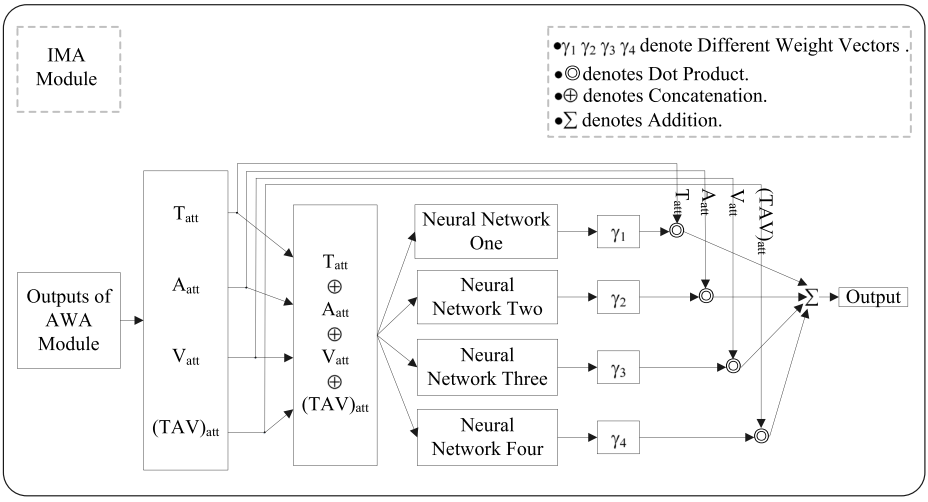

Fig. 4 Overview of IMA module

their output weight vectors as γ_1, γ_2, γ_3, and γ_4, respectively. The computations are as follows:

$$\gamma_1 = NN_1(concatAll) \tag{8}$$

$$\gamma_2 = NN_2(concatAll) \tag{9}$$

$$\gamma_3 = NN_3(concatAll) \tag{10}$$

$$\gamma_4 = NN_4(concatAll) \tag{11}$$

At each timestep t, γ_1 assigns how much of the input of T_{att} to remember, γ_2 assigns how much of the input of A_{att} to remember, γ_3 assigns how much of the input of V_{att} to remember, and γ_4 assigns how much of the input of $(TAV)_{att}$ to remember.

Furthermore, each weight vector does a dot product with its assigned input. Then, all the results of the dot product operations are merged using an addition operation. We do this as:

$$output = T_{att} \cdot \gamma_1 + A_{att} \cdot \gamma_2 + V_{att} \cdot \gamma_3 + (TAV)_{att} \cdot \gamma_4 \tag{12}$$

To this, both intra-modality and inter-modality dynamics are integrated by our model. Finally, the $output$ in the IMA module is sent to a fully connected layer, which is to reduce the dimension of the prediction.

4 Experimental methodology

In this section, we perform experiments on the MOSI dataset [49]. It should be noted that the developed model needs to find generic and person-independent features because the person varies in real-world applications. We perform person independent experiments to evaluate the performance of AWMA. Since humans express their intentions in a structured manner, there are synchronizations between intentions in language, speech, and gestures. These synchronizations constitute the relations between the three modalities (text, audio, and visual).

4.1 Dataset

The MOSI dataset involves three modalities with completely different natures: text, audio, and visual. It is a collection of 93 opinion videos of speaking about certain topics from online sharing websites. Each video consists of multiple opinion segments and each segment is annotated with the sentiment score by 5 annotators. These five annotations are averaged as the sentiment polarity. The goal of our experiments is to identify a speaker's sentiment (positive or negative) based on the segment content. For binary classification, we report F1 score and accuracy as the evaluation metrics. To ensure generalization of the model, the dataset is split into train, validation, and test sets. There is no identical speakers between sets. There are a total of 2199 segments, with 1284 segments in the train set, 229 in the validation set, and 686 in the test set. Table 1 summarizes the splits of the dataset.

Table 1 Dataset splits to ensure speaker independent learing

Dataset	Partition	Segment Count	Video Count
MOSI	train	1284	52
	valid	229	10
	test	686	31

4.2 Sequence features

For a fair comparison with the state-of-the-art method [33], we employ the identical sequence features data downloaded from their public website [1]. For each of the three modalities, the procession of the information from videos is briefly described as follows. Readers can refer to the literature [46, 47] for the details.

Text modality The MOSI dataset provides manual transcriptions. The pre-trained glove word embeddings [20] are used to convert the transcripts of videos into a sequence of word vectors. Words are considered the basic units of text and the interval duration of each word utterance [45] is a time-step. For the aim of obtaining the exact utterance timestamp of each word, P2FA [44] forced aligner is employed to perform a forced alignment.

Audio modality The software COVAREP [6] is used to extract acoustic features from the full audio clip of each segment at 100Hz to form a sequence that represents variations in tone of voice over an audio segment. These features include 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking, and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters, and maxima dispersion quotients. The average audio sequence features are calculated for each word utterance to ensure time alignment.

Visual modality The library Facet is used to extract a set of visual features from the full video segment at 30Hz to form a sequence of facial gesture measures throughout time. These features include facial action units, per-frame basic and advanced emotions, head pose, gaze tracking, and HOG features [51]. The average visual sequence features are calculated for each word utterance to ensure time alignment.

4.3 Baseline models

We compare the performance of our AWMA with the following baselines in multimodal sentiment analysis.

TFN [45] Tensor Fusion Network (TFN) captures inter-modality and intra-modality dynamic interactions by using the tensor outer product. In their model, inter-modality dynamics are modeled with a multimodal fusion approach, named Tensor Fusion, which explicitly aggregates unimodal, bimodal, and trimodal interactions. The intra-modality dynamics are modeled through three modality embedding subnetworks, for language, visual and acoustic modalities, respectively.

LMF [13] Low-rank Multimodal Fusion (LMF) is to decompose the weights into low-rank factors, which reduces the number of parameters and computational complexity. This LMF effectively improves the training and testing efficiency compared to TFN which performs multimodal fusion with tensor representations.

MARN [47] Multi-attention Recurrent Network (MARN) models interactions between modalities through time using a neural component called the multi-attention block (MAB) and storing them in the hybrid memory called the long-short term hybrid memory (LSTHM).

[1] https://github.com/A2Zadeh

Table 2 Performance of AWMA in terms of Accuracy and F1-score on the MOSI dataset

Models	Accuracy	F1
TFN	73.9	73.4
LMF	76.4	75.7
MARN	77.1	77.0
MFN	77.4	77.3
MFM	78.1	78.1
AWMA	**80.0**	**79.9**

MFN [46] Memory Fusion Network (MFN) considers to model the intra-modality and inter-modality interactions through time with a delta-memory attention network and summarize them with a multi-view gated memory.

MFM [33] Multimodal Factorization Model (MFM) is a latent variable model with conditional independence assumptions over multimodal discriminative factors and modality-specific generative factors. Multimodal discriminative factors shared across all modalities contain joint multimodal features required for discriminative tasks such as sentiment prediction. Modality-specific generative factors are unique for each modality and contain the information required for generating data. The MFM objective is derived by approximating the joint-distribution Wasserstein distance via a generalized mean-field assumption.

5 Results and discussion

We report results in classification accuracy and F1 score as a metric. Higher values denote better performance. The results of other models are from the literature [13, 33]. Table 2 summarizes the comparisons between AWMA and the baselines for sentiment analysis

Table 3 The variation of Accuracy and F1 under different $unificationSize$ values

$unificationSize$	Accuracy	F1
16	56.5	55.9
24	73.9	73.8
32	72.0	72.0
40	61.2	60.7
48	73.9	73.9
56	75.0	74.9
64	55.2	54.6
72	**80.0**	**79.9**
80	52.9	50.0
88	52.6	48.6
96	69.5	68.9
104	63.8	63.8
112	58.4	58.5
120	70.8	70.8
128	76.3	76.3

Table 4 Ablation study for attention mechanisms of AWMA on the MOSI dataset

AWMA variants	Accuracy	F1
$AWMA-T$	74.2	74.2
$AWMA-A$	73.2	73.2
$AWMA-V$	73.8	73.7
$AWMA-(T\oplus A\oplus V)$	73.8	73.8

in the MOSI dataset. As shown in Table 2, our model AWMA outperforms the compared models with classification performance increase in Accuracy ranging from 1.9% to 6.1% and in F1 score ranging from 1.8% to 6.5%. Especially, our method AWMA succeeds over the state of the art [33] by 1.9% in Accuracy. This highlights the capability of our proposed AWMA model in understanding the sentiment aspect of multimodal communication.

AWA module is the core of AWMA. As the above description, the AWA module updates the t timestamp feature X_t by adding $attC_{hf}$ in the final process (7). Then, the updated X_t after dimension unification (refer to Section 3.2) is sent to the IMA module. Set the size of dimension unification as $unificationSize$, Table 3 summarizes the variation of Accuracy and F1 under different $unificationSize$ values. From Table 3, it shows that the different values of $unificationSize$ can have a great influence on performance on the dataset MOSI. When $unificationSize$ is 72, the Accuracy and F1 are the best.

The main novelty of our AWMA model is to represent the asymmetric weights of the contexts using asymmetric windows. Hence, to understand the effects of AWA modules, we pose four variations (marked as $AWMA-T$, $AWMA-A$, $AWMA-V$, and $AWMA-(T\oplus A\oplus V)$) to conduct the ablation study. Concretely, the variation $AWMA-T$ denotes that the AWA module corresponding to aspect T is removed; the variation $AWMA-A$ denotes that the AWA module corresponding to aspect A is removed; the variation $AWMA-V$ denotes that the AWA module corresponding to aspect V is removed; the variation $AWMA-(T\oplus A\oplus V)$ denotes that the AWA module corresponding to aspect $(T\oplus A\oplus V)$ is removed.

Table 4 demonstrates the results of the ablation study. We can see that the performance of the variation $AWMA-A$ (without the AWA module corresponding to the aspect A) falls by 6.8% in Accuracy, which means that the AWA module corresponding to aspect A stands most important. Besides, the AWA module corresponding to aspect T is also important, but less than that corresponding to aspect A and aspect V as its absence causes performance to fall by 5.8%. We speculate the reason to be that the context delay on the aspect T is less than that on aspect A and V.

6 Conclusion

In this paper, we have presented a multimodal neural network named AWMA, to analyze the sentiment in user-generated videos. In contrast to the state-of-the-art method, MFM, our method considers that the historical and future contexts at a particular timestamp of the input data should be differentiated. Further, these implicit weights of the contexts can be represented with asymmetric windows. Our model AWMA can outperform the compared models with an increase of accuracy performance ranging from 1.9% to 6.1%. In the future, we search for a better combination setting of the values of the asymmetric windows and

improve the quality of multimodal feature data. At the same time, because Graph Neural Network (GNN) has its unique advantages compared with other neural networks, we will consider multimodal causability with GNN.

Acknowledgements This research was supported in part by Science and Technology Program of Guangzhou (202102020878), National Natural Science Foundation of China (62006053), Special Innovation Project of Guangdong Education Department (2018KQNCX072), the Youth Innovative Talents Project in Guangdong Universities (2020KQNCX186), the Fourth College Level Project of Guangdong Justice Police Vocational College (2020YB16), the 13th Five-Year Plan of Guangdong Institute of Higher Education Research on Higher Education of Young Teachers in Colleges and Universities in 2019 (19GGZ070), and thanks Ziang Liu for revising the english grammar of the paper.

References

1. Baecchi C, Uricchio T, Bertini M, Del Bimbo A (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed Tools Appl 75(5):2507–2525
2. Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443
3. Cavallari S, Zheng VW, Cai H, Chang C-CK, Cambria E (2017) Learning community embedding with community detection and node embedding on graphs. In: Proceedings of the 2017 ACM on conference on information and knowledge management. ACM, pp 377–386
4. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv:1406.1078
5. Datcu D, Rothkrantz LJM (2014) Semantic audio-visual data fusion for automatic emotion recognition. Emotion recognition: a pattern analysis approach 411–435
6. Degottex G, Kane J, Drugman T, Raitio T, Scherer S (2014) Covarep—a collaborative voice analysis repository for speech technologies. In: 2014 Ieee international conference on acoustics, speech and signal processing (icassp). IEEE, pp 960–964
7. Ebrahimi M, Hossein Yazdavar A, Sheth A (2017) Challenges of sentiment analysis for dynamic events. IEEE Intell Syst 32(5):70–75
8. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv:1606.01847
9. Gunes H, Piccardi M (2007) Bi-modal emotion recognition from expressive face and body gestures. J Netw Comput Appl 30(4):1334–1345
10. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computat 9(8):1735–1780
11. Hussain A, Huang G-B (2015) Towards an intelligent framework for multimodal affective data analysis. Neural Netw 63:104–116
12. Kapoor A, Burleson W, Picard WR (2007) Automatic prediction of frustration. Int J Human-Comput Stud 65(8):724–736
13. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, Morency L-P (2018) Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th annual meeting of the association for computational linguistics, pp 2247–2256
14. Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S (2018) Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowl-Based Syst 161:124–133
15. Mekhaldi D, Lalanne D, Ingold R (2012) A multimodal alignment framework for spoken documents. Multimed Tools Appl 61(2):353–388
16. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) M3er Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 1359–1367
17. Morency L-P, Mihalcea R, Doshi P (2011) Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proceedings of the 13th international conference on multimodal interfaces. ACM, pp 169–176
18. Nojavanasghari B, Gopinath D, Koushik J, Baltrušaitis T, Morency L-P (2016) Deep multimodal fusion for persuasiveness prediction. In: Proceedings of the 18th ACM international conference on multimodal interaction. ACM, pp 284–288

19. Pandeya YR, Lee J (2021) Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimed Tools Appl 80(2):2887–2905
20. Pennington J, Socher R, Manning DC (2014) Glove Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
21. Pérez-Rosas V, Mihalcea R, Morency L-P (2013) Utterance-level multimodal sentiment analysis. In: Proceedings of the 51st annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 973–982
22. Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: From unimodal analysis to multimodal fusion. Inform Fusion 37:98–125
23. Poria S, Cambria E, Gelbukh A (2015) Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 2539–2544
24. Poria S, Cambria E, Howard N, Huang G-B, Hussain A (2016) Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing 174:50–59
25. Poria S, Chaturvedi I, Cambria E, Hussain A (2016) Convolutional mkl based multimodal emotion recognition and sentiment analysis. In: 2016 IEEE 16Th international conference on data mining (ICDM). IEEE, pp 439–448
26. Pun T, Alecu TI, Chanel G, Kronegg J, Voloshynovskiy S (2006) Brain-computer interaction research at the computer vision and multimedia laboratory, University of Geneva. IEEE Trans Neural Syst Rehabilit Eng 14(2):210–213
27. Rajagopalan SS, Morency L-P, Baltrusaitis T, Goecke R (2016) Extending long short-term memory for multi-view structured learning. In: European conference on computer vision. Springer, pp 338–353
28. Ren J, Hu Y, Tai Y-W, Wang C, Xu L, Sun W, Yan Q (2016) Look, listen and learn—a multimodal lstm for speaker identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 30
29. Shan C, Gong S, McOwan PW (2007) Beyond facial expressions: Learning human emotion from body gestures. In: BMVC, pp 1–10
30. Sohrab F, Raitoharju J, Iosifidis A, Gabbouj M (2021) Multimodal subspace support vector data description. Pattern Recogn 110:107648
31. Song Y, Morency L-P, Davis R (2012) Multi-view latent variable discriminative models for action recognition. In: 2012 IEEE Conference on computer vision and pattern recognition. IEEE, pp 2120–2127
32. Song Y, Morency L-P, Davis R (2013) Action recognition by hierarchical sequence summarization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3562–3569
33. Tsai HY-H, Liang PP, Zadeh A, Morency L-P, Salakhutdinov R (2018) Learning factorized multimodal representations. arXiv:1806.06176
34. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show, Tell A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
35. Wang W, Arora R, Livescu K, Bilmes J (2015) On deep multi-view representation learning. In: International conference on machine learning. PMLR, pp 1083–1092
36. Wang H, Meghawat A, Morency L-P, Xing EP (2017) Select-additive learning: Improving generalization in multimodal sentiment analysis. In: 2017 IEEE International conference on multimedia and expo (ICME). IEEE, pp 949–954
37. Wörtwein T, Scherer S (2017) What really matters—an information gain analysis of questions and reactions in automated ptsd screenings. In: 2017 Seventh international conference on affective computing and intelligent interaction (ACII). IEEE, pp 15–20
38. Xing FZ, Cambria E, Welsch RE (2018) Natural language based financial forecasting: A survey. Artif Intell Rev 50(1):49–73
39. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, Attend and tell neural image caption generation with visual attention. In: International conference on machine learning. PMLR, pp 2048–2057
40. Xu J, Gavves E, Fernando B, Tuytelaars T (2015) Guiding the long-short term memory model for image caption generation. In: Proceedings of the IEEE international conference on computer vision, pp 2407–2415
41. Xu C, Tao D, Xu C (2013) A survey on multi-view learning. arXiv:1304.5634
42. Young T, Cambria E, Chaturvedi I, Zhou H, Biswas S, Huang M (2018) Augmenting end-to-end dialogue systems with commonsense knowledge. In: Thirty-second AAAI conference on artificial intelligence
43. Yu W, Zeiler S, Kolossa D (2021) Multimodal integration for large-vocabulary audio-visual speech recognition. In: 2020 28Th european signal processing conference (EUSIPCO). IEEE, pp 341–345
44. Yuan J, Liberman M (2008) Speaker identification on the scotus corpus. J Acoust Soc Am 123(5):3878

45. Zadeh A, Chen M, Poria S, Cambria E, Morency L-P (2017) Tensor fusion network for multimodal sentiment analysis. arXiv:1707.07250
46. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency L-P (2018) Memory fusion network for multi-view sequential learning. In: Thirty-second AAAI conference on artificial intelligence
47. Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency L-P (2018) Multi-attention recurrent network for human communication comprehension. In: Thirty-second AAAI conference on artificial intelligence
48. Zadeh A, Zellers R, Pincus E, Morency L-P (2016) Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259
49. Zadeh A, Zellers R, Pincus E, Morency L-P (2016) Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intell Syst 31(6):82–88
50. Zeng Z, Tu J, Liu M, Huang TS, Pianfetti B, Roth D, Levinson S (2007) Audio-visual affect recognition. IEEE Trans Multimed 9(2):424–428
51. Zhu Q, Yeh M-C, Cheng K-T, Avidan S (2006) Fast human detection using a cascade of histograms of oriented gradients. In: 2006 IEEE Computer society conference on computer vision and pattern recognition (CVPR'06), vol 2. IEEE, pp 1491–1498