Multimodal Sentiment Analysis Based on Multiple Stacked Attention Mechanisms

Chao He[1,2], Yingshan Shen[1], Nan Zhong[1], Dongqing Song[2], Huijuan Hu[2], Dingju Zhu[3*], Lihua Cai[1*]
[1]Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan, China
[2]School of Computer Science, South China Normal University, Guangzhou, China
[3]School of Software, South China Normal University, Foshan, China
Email:{chaohe, yshen, jonathan, 2019022679, 2021023279, zhudingju, lee.cai}@m.scnu.edu.cn

***Abstract*—Deciphering sentiments or emotions in face-to-face human interactions is an inherent capability of human intelligence, and thus a natural goal of artificial intelligence. The proliferation of multimedia data in video sites gives rise to multimodal sentiment analysis in various applications and research fields such as movie and product review, opinion polling, and affective computing. In order to improve the performance of multimodal sentiment analysis task, this paper proposes a novel neural network with multiple stacked attention mechanism (MSAM) on multimodal data containing texts, video, and audio at an utterance level. We conduct experiments using two benchmark datasets, namely CMU Multi-modal Opinion-level Sentiment Intensity (CMU-MOSI) corpus, and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) corpus. Compared with a comprehensive set of state-of-the-art baselines, the evaluation results demonstrate the effectiveness of our proposed MSAM network.**

***Index Terms*—multimodality; attention mechanism; sentiment analysis; feature fusion**

I. Introduction

The rapid development of information technology and social networks leads to the generation of vast amount of multimedia data. These diverse data make possible sentiment analysis, which aims to understand the opinions expressed by stakeholders on various entities such as products, movies, political stances and beyond [1]. It has initially focused on textual data, and is widely applied within various social media (e.g., WeChat, Twitter) and e-commerce (e.g., Amazon and Walmart) for user emotion tracking and product reviews. As the popularity of video-based social media (e.g., YouTube, Facebook, and TikTok) increases, a large amount of videos are uploaded by users to these platforms, adding extra dimensions of data for sentiment analysis [2]. Videos usually contain both spoken words that can be transcribed into texts, and audios, from which utterance features that express sentiments or emotions can be extracted, in addition to the visual images.

Multimodal sentiment analysis refers to sentiment analysis on multimodal data that usually contain text, images and audio data [3], [4]. Video is the main carrier of these multimodal data: images extracted as frames from video carry facial expressions and gestures; texts transcribed from video carry semantics; and audio embedded in video carries tone and pitch signals [5], all of which, if combined, can facilitate sentiment analysis. These three modalities complement and corroborate the identification of the underlying sentiments expressed in the video, and is thus expected to be more efficient in sentiment analysis when compared to only using textual data. In addition, viewers can distinguish multiple utterances in a video, and these utterances can have different emotional tendencies, and the emotional information of each utterance is often interdependent with the other utterances as its contexts [6]. This injects vitality into multimodal sentiment analysis, but simultaneously also poses numerous challenges.

Compared with text-based sentiment analysis, multimodal sentiment analysis faces the following challenges: 1) leveraging information from extra modalities requires feature extractions for each unique modality; 2) fusing the heterogeneous features from different modalities is a non-trivial task; and 3) filtering counter-effects among the modalities, especially when noise exists in some of them, is critical to guarantee better model performance. To address these challenges, this paper proposes a novel neural network architecture with Multiple Stacked Attention Mechanism (MSAM). Various attention layers are applied and stacked to learn the filtering capability for our network model, which enpowers it to discern modal and feature contributions.

To summarize, the contributions of this work include the following: **(1) We propose a novel network architecture with multiple stacked attention mechanism for multimodal sentiment analysis. (2) To the best of our knowledge, this is the first work to adopt utterance level attention mechanism in this increasingly popular domain of multimodal sentiment analysis. (3) Through comprehensive evaluations, we show that our proposed approach outperforms the state-of-the-art baseline algorithms in multimodal sentiment analysis.**

In the remaining presentation, we summarize the most relevant literature in Section II, illustrate the design of our proposed network architecture in Section III, explain our experiments and present the results in Section IV, followed by a discussion in Section V and a conclusion in Section VI.

II. Related Work

Multimodal sentiment analysis is a relatively new field when compared to text-based sentiment analysis [7], [8]. Zadeh et al. [9] proposed a tensor fusion network model to learn the intra- and inter-modal interactions of text, visual, and acoustics. Chen et al. [10] proposed a multimodal embedding using

gated multimodality, a model for multimodal input word-level fusion with long-short-term memory (LSTM) and temporal attention modules. Lan et al. [11] proposed an attention-based deep learning model called DGCCA-AM for multimodal emotion recognition. It realizes adaptive modality fusion through extended canonical correlation analysis method and attention mechanism. Li et al. [12] applied a multi-head attention and soft-attention modules to learn mapping relationships between multi-modalities, aiming to address modality fusion challenge in multimodal sentiment analysis. Han et al. [13] proposed an end-to-end network that is capable of fusing correlation increment while separating difference increment between multimodalities. Similar to our current work, Hao et al. [14] leveraged multi-head attention with Bi-GRU to encode multimodal data for sentiment analysis. The above works mainly focus on modal fusion, while pay less attention to investigate the dependencies between utterances. Poria et al. [15] proposed a LSTM-based framework, which exploits contextual information to capture the interdependencies between utterances.

Although the above researchers have proposed new methods that demonstrate certain improvements on multimodal sentiment analysis, they are limited in two major ways. On one hand, in learning multimodal feature representation, it is necessary to fully consider not only the representation of internal single-modal features, but also the interactions and fusions of them across modalities. On the other hand, the introduction of multimodal data creates information redundancy, so that these models cannot focus on the most important features from different modalities. This is a promising avenue to explore to further improve model performance for multimodal sentiment classification.

III. MSAM Model

To solve the challenges in intra-modal feature representation and inter-modal feature fusion, this paper proposes a novel network architecture with Multiple Stacked Attention Mechanism (MSAM) for multimodal sentiment analysis tasks. Figure 1 presents the proposed network architecture of MSAM with three modalities in a sentiment analysis task. These modalities include texts, images, and audio, and can be easily extended to other data modalities.

A. Problem Definition

Let D denotes a dataset consists of m videos, i.e., $D = [V_1, V_2, V_3, \ldots, V_m]$, where the i^{th} video consists of n utterances, i.e., $V_i = [u_{i1}, u_{i2}, u_{i3}, \ldots, u_{in}]$, $1 \leq i \leq m$, where u_{i1} is the first segment in the video and n is the total number of utterances in video V_i. For each utterance u_{ij}, $1 \leq i \leq m$, $1 \leq j \leq n$, its features can be represented as feature vectors containing three modalities $u_{ij} = [t_{ij}, v_{ij}, a_{ij}]$, where t_{ij}, v_{ij}, and a_{ij} are the textual, visual, and audio feature representations of the j^{th} utterance in the i^{th} video, respectively. Assuming a total of C sentiment categories in users' videos, the goal of our proposed approach is to label the sentiment categories of each utterance u_{ij} in each video. For this task, other utterances in a video other than utterance u_{ij} are regarded as the contexts of u_{ij}, and the accuracy is used as the performance metric of the model.

B. Unimodal Feature Extraction

To facilitate experimental comparisons with other baseline methods, our single modality feature extractor is consistent with other popular methods. For the CMU-MOSI [16] dataset, we adopt the features provided by Poria et al. [17] as the input to the MSAM model. Specifically, the convolutional neural network proposed by Karpathyet et al. [18] is used to extract text features; the 3D convolutional neural network proposed by Ji et al. [19] is used to extract visual features, and the openSMILE approach proposed by Eyben et al. [20] is used to extract speech features. For the CMU-MOSEI [21] dataset, we adopt the CMU-Multi-modal Data SDK [8] tool to extract features from different modalities. Specifically, the text, audio and video features are extracted by GloVe [22], CovaRep [23] and Facets[1] , respectively.

The above unimodal features are extracted from the respective unimodal classifiers [15], and contextual relations between utterances are not considered at this stage. In order to capture contextual information between the utterances in each modality, we feed the extracted unimodal features from each modality into a bidirectional Gated Recurrent Unit (Bi-GRU). BiGRU is composed of two GRUs in opposite directions, which can effectively capture long-term dependencies of sequntial data, significantly mitigating both the gradient vanishing and explosion problems in the training recurrent neural network [17]. In BiGRU, the forward and reverse passes of the input feature vectors will participate in the hidden layer calculation at the corresponding moment, and the output features are concatenated to obtain text, visual and audio features with contextual information.

Considering the internal heterogeneity of utterance sequences within each data modality, we try to fully exploit the within modality correlations of single-modal features. The context-dependent single-modal feature sequences obtained by the above BiGRU are mapped separately through the Dense layer in the time dimension into their respective semantic spaces. The transformation process is as follows:

$$D_{mod} = tanh(W_{mod}B_{mod} + b_{mod}) \tag{1}$$

where B_{mod}, $mod \in \{t, a, v\}$ are the output features of text, audio, and video through BiGRU, W_t, b_t, W_a, b_a, W_v, and b_v are the parameters learned in the network, $D_t \in R^{u \times d}$, $D_a \in R^{u \times d}$, $D_v \in R^{u \times d}$ represent the final output of the fully connected layer in the time dimension in texts, visuals, and speech, respectively, u represents the total number of sentences, and d represents the number of neurons in the fully connected layer.

C. Multimodal Feature Fusion

For multimodal sentiment analysis, data from different modalities contain sentimental tendencies that could be differ-

[1] https://imotions.com/biosensor/fea-facial-expression-analysis/

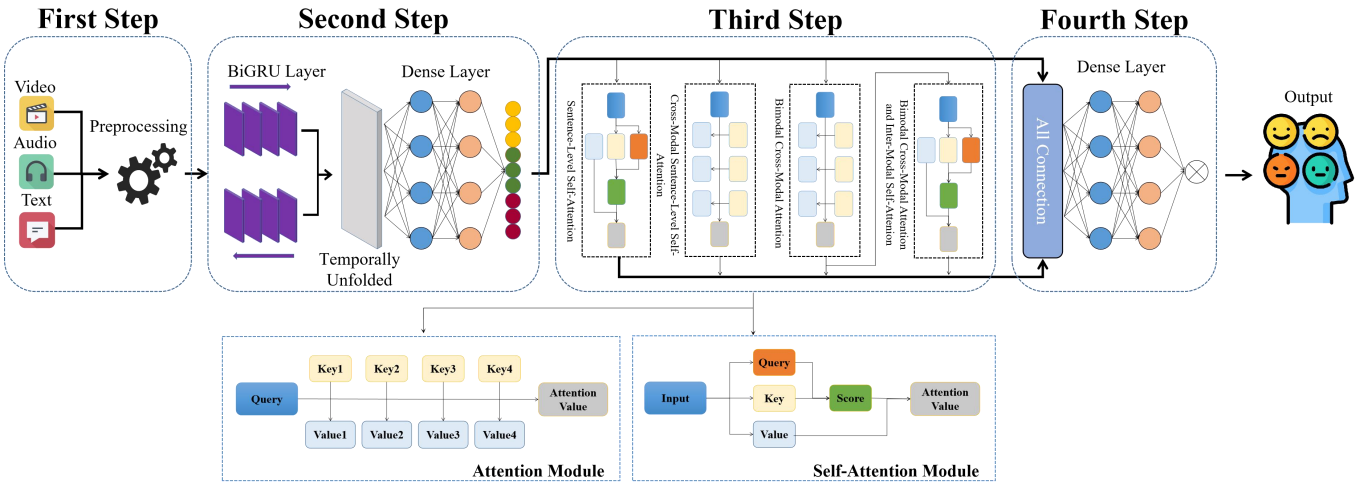

Fig. 1. MSAM Architecture. (1) First step - preprocessing input multimodal data. (2) Second step - BiGRU + Fully connected dense layer(s). (3) Third step - four attention and self-attention layers. (4) Fourth step - concatenation + fully connected dense layer(s) with softmax activation function for prediction.

ent from or complementary to each other. Therefore, depending on the internal modeling of different modalities, data from each modality could be fused with those of the other modalities differently. Complementary effects can effectively improve the performance of the model [24]. In addition, feature fusion is the core problem in multimodal research, and information from different modalities has different contributions to sentiment classification. Therefore, identifying the contribution of each modality is of critical importance to enhance model performance.

The proposed MSAM architecture in this work fuses features through different levels of self-attention and attention mechanisms. Self-attention [25] helps to capture the internal correlations of data, while reducing the dependencies on external information. Cross-modal attention is capable of capturing the correlations between different modalities, so as to encode the dynamic interactions between different modalities. When modeling multimodal emotional features, it is critical to selectively distinguish internal features within each modality and those across multi-modalities. Identifying contributions of each modal to the target task can be achieved by attention mechanisms, which enhance the weights of important features to reduce interference from irrelevant features. In order to obtain a more efficient multimodal joint feature representation, the MSAM model leverages a four-layer attention architecture on an end-to-end fashion. Without loss of generality, let us denote the dot product with $\cdot$, the Hadamard product with $\odot$, and the concatenation operation with $\oplus$ in the rest of the paper.

The first layer in the MSAM network is the Sentence-Level Self-Attention module (SL-SAtten). Take the text modality as an example. Assume that the text modality has a total of u utterances. For each utterance x_i, $1 \leq i \leq u$, the SL-SAtten module is given as below:

$$m_i = x_i \cdot x_i^T \tag{2}$$

$$n_i = softmax(m_i) \tag{3}$$

$$o_i = n_i \cdot x_i \tag{4}$$

$$a_i = o_i \odot x_i \tag{5}$$

$$T = a_1 \oplus a_2 \oplus \ldots \oplus a_u \tag{6}$$

$$O_{SL-SAtten} = T \oplus V \oplus A \tag{7}$$

where a_i is the output of the utterance x_i after self-attention. All the outputs are concatenated to generate the corresponding text feature T. The corresponding visual feature V and the corresponding audio feature A can also be obtained in the same vein, and the final output of the SL-SAtten module $O_{SL-SAtten}$ can be obtained by concatenating these text, visual and audio features.

The second layer in the MSAM network is the Cross-modal Sentence-level Self-Attention module (CS-SAtten). Assume that t_i, a_i, and v_i are the i^{th} sentence in text, vision, and audio, respectively, the CS-SAtten module is given as below:

$$x_i = t_i \oplus v_i \oplus a_i \tag{8}$$

$$m_i = x_i \cdot x_i^T \tag{9}$$

$$n_i = softmax(m_i) \tag{10}$$

$$o_i = n_i \cdot x_i \tag{11}$$

$$a_i = o_i \odot x_i \tag{12}$$

$$O_{CS-SAtten} = a_1 \oplus a_2 \oplus \ldots \oplus a_u \tag{13}$$

where a_i is the output of the cross-modal concatenated utterance x_i through cross-modal self-attention, and can be used to capture the correlations among different modals on an utterance level. $O_{CS-SAtten}$ is the output features of the CS-SAtten module.

The third layer is the Bimodal Cross-modal Attention module (BC-Atten). This module uses the fused features from two of the three modalities as conditional vectors to strengthen the connections of important features between modalities and

interactions, while weaken the association with secondary interaction features, so as to deeply explore the interactions between different modalities. Take text and visual modalities as example, and assume that D_t and D_v are the outputs of text and visual modalities from preprocessing respectively, the BC-Atten module is given as below:

$$m_1 = D_t \cdot D_v^T, \quad m_2 = D_v \cdot D_t^T \tag{14}$$

$$n_1 = softmax(m_1), \quad n_2 = softmax(m_2) \tag{15}$$

$$o_1 = n_1 \cdot D_v, \quad o_2 = n_2 \cdot D_t \tag{16}$$

$$a_1 = o_1 \odot D_t, \quad a_2 = o_2 \odot D_v \tag{17}$$

$$O_{BC-Atten}(vt) = a_1 \oplus a_2 \tag{18}$$

where $O_{BC-Atten}(vt)$ is the output features of video and text cross-modal attention.

The fourth layer is the Inter-modal Self-Attention (I-SAtten) module, which takes the output features of the BC-Atten module as input. In order to further obtain the correlations between cross-modal features, self-attention fusion is carried out through self-attention mechanism on the concatenated cross-modal features. The I-SAtten module is given as below:

$$x = O_{BC-Atten}(vt) \oplus O_{BC-Atten}(at) \tag{19}$$

$$m = x \otimes x^T \tag{20}$$

$$n = softmax(m) \tag{21}$$

$$o = n \odot x \tag{22}$$

$$O_{I-SAtten} = o \otimes x \tag{23}$$

where $O_{I-SAtten}$ is the output features of the I-SAtten module.

Finally, the obtained feature vectors from all four attention or self-attention layers are concatenated, and inter-modal interaction fusion features and the internal features of the modality, obtained by the integration of the fully connected layer, are used for sentiment classification. The calculation process is as follows:

$$\begin{aligned} O = {} & D_v \oplus D_t \oplus D_a \\ & \oplus O_{SL-SAtten} \oplus O_{CS-SAtten} \\ & \oplus O_{BC-Atten}(vt) \oplus O_{BC-Atten}(at) \\ & \oplus O_{I-SAtten} \end{aligned} \tag{24}$$

$$output = softmax(WO^T) \tag{25}$$

where O is the final output features of the MSAM model, and W is the learnable weights of the fully connected layer.

IV. Experimental results and analysis

Experiments are carried out on the CMU-MOSI [16] and CMU-MOSEI [21] datasets, and classification accuracy of the network is adopted as the performance metric for model evaluations.

TABLE I
Experimental parameter settings.

Parameter	Value
BiGRU unit	300
BiGRU Dropout	0.5
fully connected unit	100
fully connected Dropout	0.7
activation function	tanh
learning rate	0.001
batch processing	32
number of iterations	64
optimization function	Adam
loss function	categorical crossentropy

A. Datasets

(1) CMU-MOSI contains 93 videos from YouTube. These videos include topics such as movies, products, and books. There are 2199 utterances in total, and each utterance is associated with its corresponding sentiment label. We use the corresponding preprocessing methods to obtain utterance-level features dimensions of 100, 73, and 100 in text, speech, and image, respectively. The sentiment label of each utterance is marked as positive or negative. The first 62 videos are treated as the training set, and the remaining 31 videos as the test set, while 20% of the data from the training set are randomly selected as the validation set. This results in 1447 training and 752 test utterances, respectively.

(2) CMU-MOSEI includes 3229 videos with a total of 22676 utterances. Each utterance is marked with an emotional label in a $[-3, 3]$ Likert scale, and its emotional label score is in a $[0, 3]$ Likert scale for presence of emotion. We extract 300, 74, and 35 features in text, speech, and image, respectively. In order to perform sentiment classification, we treat sentiment scores greater than or equal to 0 as positive, and less than 0 as negative. 2250, 679, and 300 videos are used as the training test, and validation sets, respectively. This results in 16216 training and 4625 test utterances.

B. Experimental setup

The experimental environment in this paper is the Keras deep learning framework in Python programming language based on TensorFlow [26]. We use a NVIDIA GeForce RTX 2060 GPU with an 8G memory. Table I summarizes the hyper-parameters for the experiments.

C. Baseline Models

We adopt the following baseline methods on multimodal sentiment analysis: GME-LSTM [10], MARN [8], TFN [9], DialogueRNN [27], BC-LSTM [7], Multilogue-Net [28], ConBIAM [24], AMF-BiGRU [29], MMMU-BA [6], MAM [5], CIM-Att [30], Graph-MFN [21], MFN [31].

D. Results

From Table II, we observe that our proposed MSAM network outperforms the baseline models on both the CMU-MOSI and CMU-MOSEI datasets with more than 1% in classification accuracy. In order to further analyze the impacts

TABLE II
COMPARISON OF PERFORMANCE ON DIFFERENT MODELS.

Network Model	Accuracy (%)	
	CMU-MOSI	CMU-MOSEI
GME-LSTM [10]	76.50	NA
MARN [8]	77.10	NA
TFN [9]	77.10	NA
DialogueRNN [27]	79.80	NA
BC-LSTM [7]	80.30	NA
Multilogue-Net [28]	81.19	NA
Con-BIAM [24]	81.91	NA
AMF-BiGRU [29]	82.05	NA
MAM [5]	81.00	81.00
MMMU-BA [6]	82.31	79.80
CIM-Att [30]	NA	80.50
MFN [31]	NA	77.70
Graph-MFN [21]	NA	76.90
MSAM	**83.45**	**82.10**

TABLE III
COMPARISON OF DIFFERENT MODALITY COMBINATIONS ON CMU-MOSI DATASET.

Modality	T+A	V+T	A+V	A+V+T
Dialogue-RNN [27]	79.80	78.90	73.90	79.80
Multilogue-Net [28]	80.18	80.06	75.16	81.19
MMMU$_B A1001$[6]	80.05	80.27	**80.29**	82.31
Con$_B IAM1001$[24]	80.45	80.98	63.96	81.91
MSAM	**81.20**	**81.98**	66.63	**83.31**

of different modality combinations on classification accuracy, we select four of the baseline methods and our proposed network on four modality combinations on the CMU-MOSI dataset. Table III shows the experimental results. T, A, and V are shorthands for text, audio, and video. We observe that when the textual modality is included, our proposed network outperforms the baseline approaches. This infers that the MSAM network relies the most on the text modality. Also, using only the audio and video (A+V) modality leads to performance degradation in all approaches, which is consistent with the results in [25]. This suggests that the polarity of emotional expression extracted from speech and video modalities is weaker than the one from text data. It could be caused by greater noise interference in these signal channels. Last but not least, the best performance is achieved when all three modalities are adopted, which corroborates the advantages for multimodal data on sentiment analysis.

E. Ablation experiment

The core of the MSAM network lies in the utterance level attention layers. In order to further analyze the contribution of the joint features from each attention module proposed in the MSAM model to the final classification accuracy, an ablation experiment was conducted on the CMU-MOSI dataset.

All the experiment results are shown in Table IV. Specifically, MSAM(BiGRU) is the classification performance of text, audio and video features through BiGRU and fully connected layers. MSAM(SL-SAtten) is based on MSAM(BiGRU) while adding the sentence-level self-attention layer. In the same vein, MSAM(CS-Atten) is based on MSAM(SL-SAtten) while adding cross-modal sentence-level self-attention layer; MSAM(BC-Atten) is based on MSAM(CS-Atten) while adding the bimodal cross-modal attention layer; and MSAM(I-SAtten) is based on MSAM(CS-Atten) while adding the inter-modal self-attention layer. It can be seen that each attention module of the MSAM model in this paper improves the performance of multimodal sentiment classification task to a marginal degree, while the best performance is achieved by the full MSAM network.

TABLE IV
COMPARISON OF THE MODELS WITH DIFFERENT ATTENTION LAYERS.

Attention Module	Accuracy(%)
MSAM(BiGRU)	80.85
MSAM(SL-SAtten)	81.25
MSAM(CS-Atten)	81.52
MSAM(BC-Atten)	81.78
MSAM(I-SAtten)	81.91
MSAM	**83.45**

V. DISCUSSION

Modal fusion is an important research direction in application areas that involve multimodal data sources. Specifically, automatic sentiment detection in videos is an increasingly important task as these data are generated by social media platforms and news organizations with unprecedented speed and volume. Sentiment signals contained in different channels (e.g., text, image, audio) of a video carry complicated patterns that could either corroborate or offset each other's sentiment signature. How these modalities are fused to identify the underlying sentiments expressed by the associated entity remains an untackled issue. This paper proposes an end-to-end novel network architecture with various attention mechanisms to solve multimedia data fusion challenges in sentiment analysis.

We formulate the multimodal sentiment analysis task on video data on an utterance level, and introduce four different attention layers that employed different modal fusion strategies based on intuition. The sentence level self-attention layer encodes the intra-modality patterns through temporal relationships of words within an utterance. In the same vein, we leverage the parallel relationship of the three different modalities by concatenating them first on a sentence level, and applied similar self-attention operator on them to encode the cross-modal patterns through temporal relationships of 'tri-words' within an utterance. Next, we introduce the bimodal cross-modal self-attention mechanism, replacing the 'word' level fusion within each utterance with utterance level fusion on two different modality combinations, text+video and text+audio. Lastly, the output from the bimodal cross-modal attention layer serves as the input for another attention layer to encode the interaction patterns from all three modalities. This systematic fusion strategy could effectively capture as much as possible relevant patterns that can identify the underlying sentiments in the utterance level of videos, leading to a better algorithm as in MSAM network. However, the introduction of

multiple attention layers leads to heavier model training burden and longer execution time. Computation efficiency in modern edge computing devices is critical, therefore we plan to design lighter algorithm based on the ideology of the current work. Also, given the tremendous efforts in obtaining label data, a modified semi-supervised version of our current approach could be another direction of our future endeavors.

VI. Conclusion

In this work, we propose a novel network with multiple stacked attentions for multimodal sentiment analysis tasks. Our approach is an end-to-end network with all incorporated attention layers being applied on an utterance level. Using two benchmark datasets published by CMU researchers on multimodal sentiment analysis, we are able to show that, when compared with a comprehensive set of baseline approaches, our proposed network achieves the best classification accuracy. This could be attributed to the capability of the various attention layers in filtering out noisy features, and the incorporation of bimodal and cross-modal features in the network architecture.

References

[1] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer, 2017.

[2] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[3] Sun Yingying, Jia Zhentang, and Zhu Haoyu. A survey of multimodal deep learning. volume 56, pages 1–10, 2020.

[4] Mahesh G Huddar, Sanjeev S Sannakki, and Vijay S Rajpurohit. A survey of computational approaches and challenges in multimodal sentiment analysis. *Int. J. Comput. Sci. Eng*, 7(1):876–883, 2019.

[5] Song Yunfeng, Ren Ge, Yang Yong, and Fan Xiaochao. Attention-based multi-level hybrid fusion for multi-task and multi-modal sentiment analysis. 2022.

[6] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3454–3466, 2018.

[7] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883, 2017.

[8] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[9] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[10] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 163–171, 2017.

[11] Yu-Ting Lan, Wei Liu, and Bao-Liang Lu. Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2020.

[12] Zuhe Li, Qingbing Guo, Chengyao Feng, Lujuan Deng, Qiuwen Zhang, Jianwei Zhang, Fengqin Wang, and Qian Sun. Multimodal sentiment analysis based on interactive transformer and soft mapping. *Wireless Communications and Mobile Computing*, 2022, 2022.

[13] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 6–15, 2021.

[14] Hao Ai, Ying Liu, and Jie Fang. Multi-head attention with disagreement regularization for multimodal sentiment analysis. In *2021 4th International Conference on Artificial Intelligence and Pattern Recognition*, pages 561–566, 2021.

[15] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038. IEEE, 2017.

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

[17] Soujanya Poria, Haiyun Peng, Amir Hussain, Newton Howard, and Erik Cambria. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing*, 261:217–230, 2017.

[18] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[19] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[20] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[21] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[23] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014.

[24] Bao Guangbin, Li Gangle, and Wang Guoxiong. Bimodal interactive attention for multimodal sentiment analysis. volume 16, page 909, 2022.

[25] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[26] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[27] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825, 2019.

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

[29] Lai Xuemei, Tang Hong, Chen Hongyu, and Li Shanshan. Attention-based feature fusion-multimodal sentiment analysis of bidirectional gated recurrent units. *Computer Application*, 41(5):1268, 2021.

[30] Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812*, 2019.

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