

Received September 17, 2021, accepted November 2, 2021, date of publication November 8, 2021, date of current version December 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3126782

Targeted Aspect-Based Multimodal Sentiment Analysis: An Attention Capsule Extraction and Multi-Head Fusion Network

DONGHONG GU¹, JIAQIAN WANG¹, SHAOHUA CAI², CHI YANG¹, ZHENGXIN SONG¹,
HAOLIANG ZHAO¹, LUWEI XIAO¹, (Member, IEEE),
AND HUA WANG³, (Senior Member, IEEE)

¹School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 560001, China

²Center for Faculty Development, South China Normal University, Guangzhou 560001, China

³Institute of Sustainable Industries and Liveable Cities, Victoria University, Melbourne, VIC 8001, Australia

Corresponding author: Shaohua Cai (caishaohua0605@163.com)

This work was supported in part by the National Statistical Science Research Project of China under Grant 2016LY98, in part by the Characteristic Innovation Projects of Guangdong Colleges and Universities under Grant 2018KTSCX049, and in part by the Science and Technology Plan Project of Guangzhou under Grant 202102080258 and Grant 201903010013.

ABSTRACT Multimodal sentiment analysis has currently identified its significance in a variety of domains. For the purpose of sentiment analysis, different aspects of distinguishing modalities, which correspond to one target, are processed and analyzed. In this work, the researchers propose the targeted aspect-based multimodal sentiment analysis (TABMSA) for the first time. Furthermore, an attention capsule extraction and multi-head fusion network (EF-Net) on the task of TABMSA is devised. The multi-head attention (MHA) based network and the ResNet-152 are employed to deal with texts and images, respectively. The integration of MHA and capsule network aims to capture the interaction among the multimodal inputs. In addition to the targeted aspect, the information from the context and the image is also incorporated for sentiment delivered. The researchers evaluate the proposed model on two manually annotated datasets. The experimental results demonstrate the effectiveness of our proposed model for this new task.

INDEX TERMS Multimodal sentiment analysis, textual and visual modalities, feature extraction, multi-modality fusion.

I. INTRODUCTION


Sentiment analysis, also referred to as sentiment classification, aims to extract opinions from a large number of unstructured texts and classifying them into sentiment polarities, positive, neutral or negative [1]. To date, much of the work on sentiment analysis focuses on textual data [2]. Notably, with the advances of social media, it is significant to precisely capture the sentiment via the presence of different modalities (i.e., textual, acoustic and visual) [3], [4]. Recent initiatives reveal that nearly 40% of reviews on cellphone in ZOL.com contain both text and image, which attract over 3 times the attention than the text-only reviews [2]. As such, the ability to analyze sentiment on multimodal data is most pronounced.

The associate editor coordinating the review of this manuscript and approving it for publication was Patrizia Grifoni¹.

On current shopping and social platforms, seeing that the text and image information is taken to mutually reinforce and complement each other, models are dedicatedly devised to classify the sentiment polarity by using both kinds of data and their latent relation [5]. Recent publications report their achievements on the task of multimodal sentiment analysis. Xu *et al.* propose a Multi-Interactive Memory Network, together with an aspect-level multimodal sentiment analysis (ABMSA) dataset for model evaluation [2]. Yu *et al.* develop methods for target-oriented multimodal sentiment classification (TMSC) [5], [6] by integrating the attention mechanisms and the pre-trained ResNet [7]. Experimental results show that an even higher accuracy can be obtained by incorporating the image into classical sentiment analysis.

On the other hand, sentiments towards different aspects of more than one entity are discussed in the same unit of text in many scenarios. targeted aspect-based sentiment

TABLE 1. An example for TABMSA.

Image	Text	Target	Aspect	polarity
	What do health heroes look like? Dr Lucille Corti died AIDS 1996, Dr Lukwiya died Ebola 2000. 52 years serving#Uganda	Dr Lucille Corti	event	negative
			appearance	positive
		Dr Lukwiya	event	negative
			appearance	positive
		Uganda	place	neutral

analysis (TABSA) combines the challenges and the superiorities of aspect-based sentiment analysis and target-oriented sentiment analysis, and paves a way for greater depth of analysis. Namely, this task requires the detection of the aspect category and the sentiment polarity for a given targeted entity. According to Saeidi *et al.* [8], TABSA caters for more generic text by making fewer assumptions with a more delicate understanding, which is both creative and practical for sentiment analysis.

In this work, the researchers introduce a new task, namely Targeted aspect-based multimodal sentiment analysis (TABMSA), which indicates the integrating of multimodal information into TABSA to facilitate the sentiment analysis. That is, by exploiting information from texts and images, the sentiment classification result with a higher accuracy can be obtained. As illustrated in Table 1, there are three targets in the text: ‘Dr Lucille Corti’, ‘Dr Lukwiya’ and ‘Uganda’. For targets ‘Dr Lucille Corti’ and ‘Dr Lukwiya’, the aspects contain ‘event’ and ‘appearance’. Notably, the sentiment polarity for ‘appearance’ is positive according to the image while that for ‘event’ is negative according to the text. On this occasion, an approach to precisely capture the information of both texts and images is highlighted.

In this paper, the researchers propose an Attention Capsule Extraction and Multi-head Fusion network (EF-Net) on the task of TABMSA. In our model, a bidirectional-GRU and multi-head self-attention mechanism is established for text semantic information encoding while the ResNet-152 model and capsule network is employed for dealing with the image, which can maintain more related information. For multimodal interaction and fusion, the multi-head attention network is applied to maximize the contribution of each modality to sentiment delivering. Lastly, the multimodal representation, concatenating with the original semantic representation, is fed into the sentiment classifier.

This model has three main characteristics. Firstly, the researchers use multi-head self-attention network to extract the features of context. Compared with previous methods that based on bidirectional LSTM model, multi-head self-attention network can avoid the long-distance dependence problem. Secondly, in order to capture sentiment information that may be contained in the image, such as human smiling face, we use capsule network for feature extraction in visual modality, since capsule network is better at this. Finally,

unlike the previous approach of bilinear pooling, we use multi-head attention network for multimodal feature fusion, because the multi-head attention mechanism can focus on the interaction of textual and visual modality in different facet, This helps the model to capture more inter-modality correlation information. Experiments are conducted on two manually annotated multimodal datasets [5], which aims to verify the effectiveness of EF-Net comparing to the baseline method. For research purpose, our code and the annotated dataset will be released via this link: <https://github.com/Aa-dh123/EF-Net>.

II. RELATED WORK

A. TABSA

Sentiment analysis [9] and aspect-based sentiment analysis [10]–[12] cannot consider the situation that the sentence may contain multiple targets and multiple aspects. TABSA, as a more challenging sentiment analysis task, is introduced to accurately determining the sentiment polarity associated with a specific aspect of a certain target [8]. This is more in line with most realistic scenarios. TABSA needs to extract important sentiment features and perform aspect-based sentiment analysis on each target-aspect pair. As an example, to capture the important context features from both the target level and the sentence level, Ma *et al.* develop an attention-based LSTM that utilizes the commonsense knowledge of sentiment-related concepts proposed in SenticNet for external knowledge incorporating [13]. Besides, a recurrent entity network is designed and deployed to track and update entity states at the right time via word-level information and sentence-level hidden memory [14]. Liu *et al.* also proposed a recurrent entity network that applied external “memory chains” with a delayed memory update mechanism to better capture the linguistic structure [15]. To improve the ability to capture context features, the pre-training model, such as EMLo and Bert, were introduced to TABSA too. For example, Hong *et al.* convert the input representation of BERT into a sentence-pair classification task via adding a self-attention mechanism [16].

B. ABMSA

There are many application scenarios for multi-modal tasks [17], [18], and aspect-based multi-modal sentiment analysis belongs to a fine-grained multi-modal sentiment

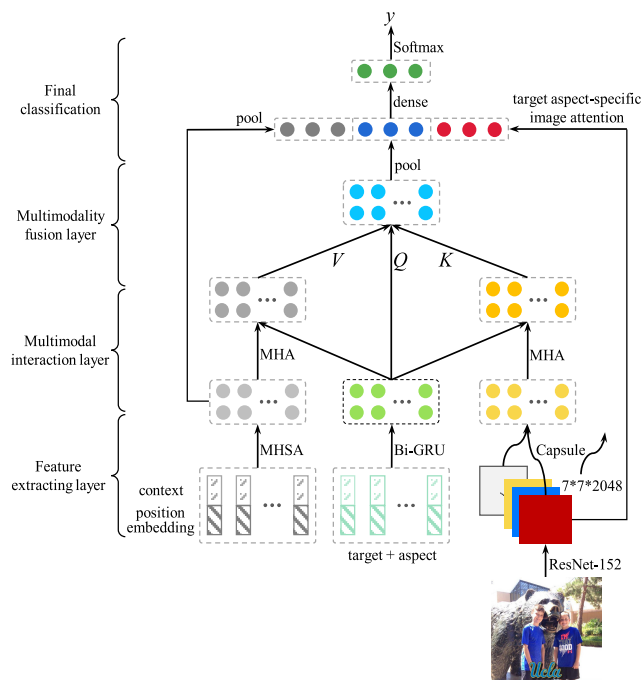


FIGURE 1. Overall architecture of the proposed EF-Net. The model first extracts the features from texts and images and encodes them into corresponding representations. Then, two multi-head attention network is carried out for multimodal information interaction. In the multimodality fusion layer, the multi-head attention-based fusion network is performed to filter and fuse the inter-modal information. The multimodality fusion outcome, together with the original semantic sequences, is concatenated for final sentiment prediction.

classification task [19]. Compared with aspect-based sentiment analysis that based on plain text, aspect-based multimodal sentiment analysis focuses on capturing sentiment features from different modal information such as textual and vision. The joint use of these modalities can not only enhance the sentiment expressing, but also improves the classification accuracy in sentiment analysis. As presented in [4], a co-memory attentional mechanism to interactively model the interaction between text and image is established, and thus to analyze the effects on one modality to the other. Motivated by the fine-grained sentiment analysis, the Multi-Interactive Memory Network is proposed to learn the interactive influences between cross-modality data and the self-influences in single-modality data [2]. Yu *et al.* proposed a Multimodal BERT architecture, which adapts BERT for cross-modal interaction to obtain target-sensitive textual/visual representations and utilize stacked multiple self-attention layers to achieve multi-modal fusion[5]. ESAFN divided the text into left context, target words and the right context, and the bilinear pooling mechanism is used for multi-modal fusion [6].

III. METHODOLOGY

The task of TABMSA can be formulated as follows: given an image I and a text sequence of n words $v^c = v_1^c, v_2^c, \dots, v_n^c$, containing m targets $v^t = v_1^t, v_2^t, \dots, v_m^t$, to characterize the aspect a . Our purpose is to figure out the sentiment polarity y towards (v^t, a) in (v^t, I) .

The architecture of EF-Net model is shown in Figure 1. Our model mainly contains four layers: feature extracting layer, multimodal interaction layer, multimodality fusion layer and final classification layer. The model firstly extracts the features from texts and images and encodes them into corresponding representations in the feature extracting layer. Then multimodal information interaction is carried out to preserve the more-related information. In the multimodality fusion layer, the multi-head attention-based fusion network is performed to filter and fuse the inter-modal information. The multimodality fusion outcome, together with the original semantic sequences, is concatenated for final sentiment prediction. The details of each part are described as follows. To start with, Multi-Head Attention (MHA) network, which is applied to our model, will be introduced briefly.

A. MULTI-HEAD ATTENTION (MHA) NETWORK

The Multi-Head Attention (MHA) aims to perform multiple attention function in parallel, which can be considered as an improvement of the traditional attention mechanism. Basically, a traditional attention is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (1)$$

where Q stands for Query, K for Key and V for Value. The regulator $\sqrt{d_k}$ is taken to constrain the dot product value. In MHA, the inputs Q, K and V are mapped through the parameter matrices. Then the attention function is computed in parallel, whose outcomes are concatenated to obtain the multi-head attention value.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \\ MultiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_n) \quad (2)$$

and where W_i^Q, W_i^K and W_i^V represent the projections parameter matrices for the corresponding inputs and $head_i$ is the attention of i -th head. In this work, the Multi-head Self-Attention (MHSA) is also employed, which can be regard as a special kind of MHA. In MHSA, the identical inputs are sent to the model, i.e., $Q = K = V$. In this way, the attention is delivered as:

$$MHSA = MultiHead(X, X, X) \quad (3)$$

where X indicates a general input of the MHA network.

B. FEATURE EXTRACTING LAYER

In this layer, both the texts and the images are sent to the model as the inputs. For the textual data, each word is mapped into a low-dimensional vector by looking up in the pretrained Glove [20]. Thus, the word embeddings for the given text are obtained. Let $w^c = \{w_1^c, w_2^c, \dots, w_n^c\}$ be the word embedding of the context, $w^t = \{w_1^t, w_2^t, \dots, w_n^t\}$ be that of the target and w^a be that of the aspect.

1) CONTEXT REPRESENTATION

In addition to the semantic information, the position information is also considered. The relative distances between each word and the target are computed and the outcomes are presented as the position embeddings $p^c = \{p_1^c, p_2^c, \dots, p_n^c\}$. By employing the MHSA mechanism, the concatenation of word embeddings and the position embeddings is transferred to the context $h^c = \{h_1^c, h_2^c, \dots, h_n^c\}$, which is delivered as:

$$h^c = MHSA([w^c, p^c]) = MultiHead([w^c, p^c], [w^c, p^c], [w^c, p^c]) \quad (4)$$

The context Representation retains the original semantic information and syntactic structure of the context to the greatest extent, so the average pooling of h^c , denoted as h_{avg}^c is used for feature fusion in the final sentiment classification:

$$h_{avg}^c = \sum_{i=1}^n h_i^c / n \quad (5)$$

2) TARGETED ASPECT REPRESENTATION

Seeing that both target and aspect are short-sequence text, Bi-GRU is employed to capture the semantic information. That is, the word embeddings of target w^t and aspect w^a are concatenated and sent to Bi-GRU network. The targeted aspect representation $h^{ta} = \{h_1^{ta}, h_2^{ta}, \dots, h_m^{ta}\}$ is given by:

$$\vec{h}_j = \vec{GRU}([w_j^t, w^a]), \quad j \in [1, m] \quad (6)$$

$$\overleftarrow{h}_j = \overleftarrow{GRU}([w_j^t, w^a]), \quad j \in [m, 1] \quad (7)$$

$$h^{ta} = [\vec{h}_j, \overleftarrow{h}_j], \quad j \in [1, m] \quad (8)$$

3) VISUAL REPRESENTATION

In order to make full use of image information, a most effective image recognition method, ResNet-152, is used for image feature extraction. For a specific input image I , first re-size it to a 224×224 -pixel image I' . With the pre-trained ResNet-152, the image feature vector R can be:

$$R = ResNet(I') \quad (9)$$

where R is a $7*7*2048$ -dimensional tensor.

Nevertheless, since *ResNet* is absent of tackling the position information of target in the image, R is fed into a one-layer capsule network afterward. Thereby, the image representation h^i , which contains the position information of the target, is written as

$$h^i = Capsule(R) \quad (10)$$

4) TARGETED ASPECT SPECIFIC IMAGE ATTENTION

Aiming to remove the unrelated context to the target (e.g., image background) and preserve the most related part, the attention mechanism is applied, based on which the more essential image representation h_{att}^i is defined as:

$$h_{att}^i = Attention(h_{avg}^{ta} W^{ta}, RW^R, RW^R) \quad (11)$$

where W^{ta} and W^R are trainable parameter matrices for mapping h_{avg}^{ta} and R into sub-space of the same dimension.

C. MULTIMODAL INTERACTION LAYER

The multimodal interaction layer is responsible for analyzing the relation between the targeted aspect, context and the image, respectively, and thus distilling the key information from the multimodal inputs. The main purpose of this layer is to obtain the targeted aspect specific textual attention and the targeted aspect specific visual attention. Therefore, MHA network is utilized to understand the interaction with respect to the targeted aspect.

For the targeted aspect h^{ta} and the context h^c , the researchers set h^{ta} as Q and h^c as K . With $K = V$, the interaction between the targeted aspect and its context is characterized by:

$$h^{tac} = MultiHead(Q, K, V) = MultiHead(h^{ta}, h^c, h^c) \quad (12)$$

where h^{tac} is the targeted aspect specific context representation.

Likewise, the representation of targeted aspect specific image h^{tai} , indicating the interaction of between the targeted aspect and the image is

$$h^{tai} = MultiHead(Q, K, V) = MultiHead(h^{ta}, h^i, h^i) \quad (13)$$

D. MULTIMODALITY FUSION LAYER

In practical use, not only the targeted aspect, but also the context and the image contain the sentiment information for determining sentiment polarities. Accordingly, following the multimodal interaction layer, the targeted aspect-specific representations from different modalities are incorporated. Instead of using a gated mechanism to control the contribution of each component, this work tends to take the three representations as the inputs of MHA model for information fusion. By exploiting the MHA mechanism, the multimodal representation is then given as:

$$h^{taci} = MultiHead(Q, K, V) = MultiHead(h^{ta}, h^{tai}, h^{tac}) \quad (14)$$

where h^{taci} stands for the multimodal representation. In eqn.(14), The researchers set h^{ta} as Q , h^{tai} as K and h^{tac} as V .

The average pooling h_{avg}^{taci} of the multimodal representation h^{taci} is get, which is further enriched by concatenating the average representation of the context and the image (h_{avg}^c and h_{att}^i).

$$h_{avg}^{taci} = \sum_{i=1}^n h_i^{taci} / n \quad (15)$$

$$O = [h_{avg}^c, h_{avg}^{taci}, h_{att}^i] \quad (16)$$

where O is the final representation with multimodal information.

E. FINAL CLASSIFICATION

The aforementioned final representation is fed into Softmax classifier for sentiment polarity distribution

identification, which is

$$x = W_O^T O + b_O \quad (17)$$

$$\begin{aligned} \tilde{y} &= \text{softmax}(x) \\ &= \frac{\exp(x)}{\sum_{k=1}^c \exp(x)} \end{aligned} \quad (18)$$

where W_O^T and b_O is the trainable weight matrix and offset vector, and C is the number of sentiment polarities.

F. MODEL TRAINING

The training process is conducted on by using the categorical cross-entropy, which is expressed as:

$$L = - \sum_i^m \sum_j^C y_i^j \log(\tilde{y}_i^j) + \lambda \| \theta \|^2 \quad (19)$$

where m is the number of aspect terms in the sentence, C is the number of sentiment polarities. The parameter y_i stands for the real sentiment distribution of i -th aspect term and \tilde{y}_i^j is the predicted one on j -th sentiment polarity. Besides, λ is the weight of L_2 regularization.

IV. EXPERIMENT

A. DATASET

In order to evaluate the performance of the EF-Net model, a large-scale TABMSA dataset is manually annotated based on two publicly available TMSA datasets Twitter15 and Twitter17 [5]. Three experienced researchers, who work on natural language processing (NLP), are invited to extract targets and aspects in the sentences and label their sentiment polarities. To start with, 500 samples from dataset are randomly picked in advance to reveal the most appearing target and aspect types, which are ‘people’, ‘place’, ‘time’, ‘organization’ and ‘other’. The targets, as well as the corresponding aspects, are presented in Table 2. In such manner, the annotated Twitter15 contains 3259 samples for training, 1148 for validation and 1059 for testing while the corresponding numbers in Twitter 17 are 3856, 1240 and 1331. Considering the TABMSA task, each sample from our dataset is composed of images and texts sentiment polarities. The expressed sentiment polarities are predefined as positive, neutral or negative. Details of our dataset is exhibited in Table 3.

B. EXPERIMENTAL SETTING

As mentioned above, experiments are conducted on dedicatedly-annotated datasets for working performance evaluation. The maximum padding length of textual content is set as 36 for Twitter15 and 31 for Twitter17. The images are sent to pre-trained ResNet-152 to obtain the $7*7*2048$ -dimension visual feature vector. For our model, the learning rate is set as 0.001, the dropout rate as 0.3 and the batch size as 128. The attention head number is 4.

TABLE 2. Targets and their respective aspects.

Target	Aspect	Target	Aspect
people	general	place	general
	event		appearance
	phenomenon		achievement
	environment		event
	experience		speech
organization	other	time	general
	general		event
	event	other	other
	other		other

TABLE 3. Statistics of Twitter15 and Twitter17 dataset.

Attribute	Twitter15	Twitter17
#Sentence	3502	2910
#Label	3	3
#Target aspect pair	5466	6427
Avg. of #Aspect/Sentence	1.6	2.2
Avg. text length/ Sentence	13.2	13.9
Max text length/ Sentence	36	31
Min text length/ Sentence	1	3

C. MODEL COMPARISON

In order to verify the superiority of our model, The researchers separately compare our model with classical textual sentiment analysis methods (ATAE-LSTM, MemNet, IAN and MGAN) and the representative multimodal sentiment analysis methods (Res-MemNet, Res-IAN, Res-MGAN, ESAFN).

ATAE-LSTM [1] applies LSTM and concatenating process to get the aspect embeddings while the attention network aims to select the word of sentiment significance.

MemNet [21] applies a multi-layer attention mechanism on top of the common word embedding layer.

IAN [22] model the representations on the foundation of the LSTM based interactive attention networks. And hidden states are taken to compute the attention scores by the pooling process.

MGAN[15] leverage a multi-grained attention network to capture the target and the context interaction at fine and course granularity.

Res-MemNet, **Res-IAN** and **Res-MGAN** take the max-pooling layer of ResNet and the hidden representation of MemNet or IAN or MGAN to concatenate for multimodality sentiment classification. Notably, for all the aforementioned model, the sentiment polarity distribution of the target is finally determined by using the Softmax classifier.

ESFAN [6] sub-divide the contexts into left context and right context based on their position to the aspect. Its hidden states are obtained via LSTM and text-level features fused via bilinear fusion layer. The visual features from images are processed using ResNet-152. Then the aspect-sensitive attention layer and the gate mechanism are employed to obtain the semantic representation of images. The multimodal feature fusion is performed by the bilinear fusion layer for sentiment classification.

TABLE 4. Comparative results of EF-Net and baselines.

	Method	Twitter15		Twitter17	
		ACC	F1	ACC	F1
Text	ATAE	68.17	59.09	65.28	62.89
	MemNet	69.49	63.10	65.96	62.80
	IAN	70.44	63.64	63.64	59.74
	MGAN	70.20	63.86	64.91	61.01
	EF-Net (Text)	71.67	67.30	66.19	62.21
Text + Visual	Res-MemNet	64.30	56.64	62.58	59.32
	Res-IAN	64.11	57.87	60.70	55.91
	Res-MGAN	70.06	61.13	64.36	61.04
	ESAFN	72.04	63.06	65.21	60.01
	EF-Net	73.65	67.90	67.77	65.32

D. MAIN RESULTS

In this experiment, the accuracy and Macro-F1 are adopted as evaluation metrics to denote the working performance. Table 4 shows the main results. In the classical TABSA tasks, it can draw the following conclusions: first of all, The ATAE model gets the worst performance, which indicates the importance of modeling the interactive relationship between context and target aspect. The ATAE model only concatenates the context representation with the aspect embedding, which was not enough to capture the correlation between them. Then, compared to MemNet, IAN and MGAN, our proposed model, which removes the image processing part, labeled as ‘EF-Net (Text)’, achieve almost the best results on two datasets. The minimum performance gap of 1.47% and 1.28% for Twitter15 and Twitter17 can be observed in Table 4 against the MGAN method. which indicates that: 1) MHSA has a better ability to capture context features than LSTM, because LSTM usually has a certain degree of long-distance dependence, while context is a longer text than target and aspect. 2) Compared with the standard attention mechanism, MHA can capture the interaction characteristics of context, target and aspect at different facets, which led to a greater ability to model interactive information. 3) The EF-Net (Text) makes use of the position information, which plays an important role in determining the sentiment expression that most relevant to target and aspect in the context. In this way, the representations in our model are considerably more informative for delivering sentiment and get a more convincing result.

On the other hand, the multimodal sentiment analysis models are generally more competitive than the basic textual sentiment analysis ones. With the integration of visual context, an even higher classification accuracy is thus accessible. On the task of TABMSA, EF-Net still significantly outperforms the baseline models. The minimum performance gap of 1.61% and 2.56% for Twitter15 and Twitter17 can be observed in Table 4 against the ESAFN method. Clearly, our model is a better alternative for the task of multimodality sentiment analysis. In spite the effectiveness of EF-Net (Text), another explanation is that the image data is fused into the texts, together with investigating the multimodal fusion with MHA network, which exploits the sentiment information and the relation of multimodalities. Since EF-Net is more capable of dealing with the TABMSA tasks, it is reasonable to expect

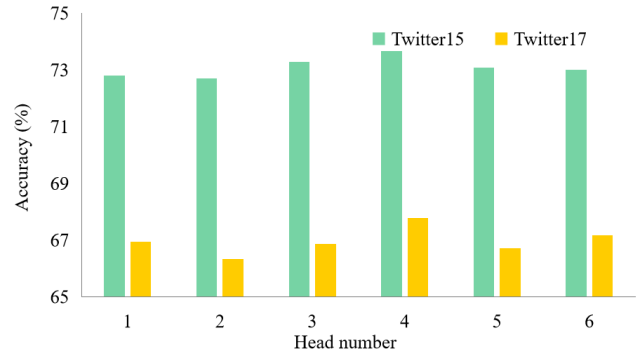


FIGURE 2. Results of difference head number.

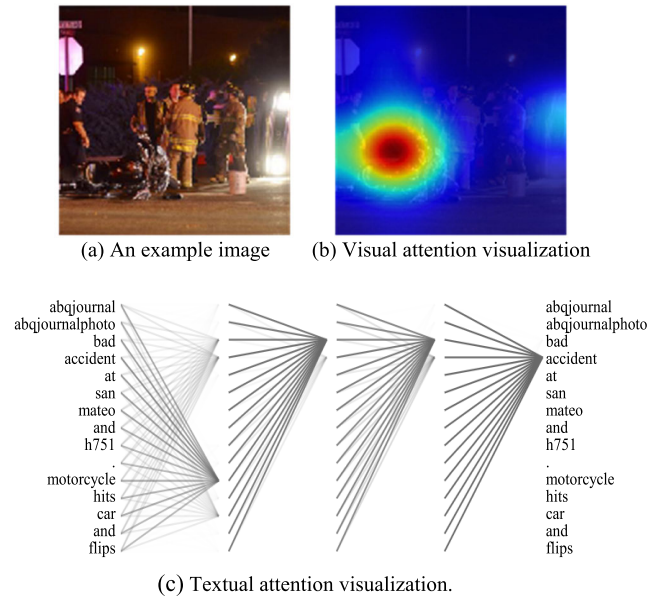


FIGURE 3. An example of visual and textual attention.

even higher accuracy in more evaluation settings, as it is the case.

E. MHA HEAD NUMBER IMPACT ANALYSIS

On this occasion, the head number of MHA is studied for better obtain the relation among modalities. At this stage, the researchers vary the number of head within the collection {1, 2, 3, 4, 5, 6}. The results on Twitter 15 and Twitter 17 of different head numbers are shown in Figure 2. One can see that our model has the highest accuracy on head number 4. For a smaller head number (i.e., 1,2,3), MHA fails to maintain all the key information especially for long texts. With the parameter increasing and the model overfitting problem, the classification accuracy drops with the head number continues increasing.

V. CASE STUDY

An example of visual and textual attention visualization is presented. Based on the MHA mechanism, both the aspect specific-contexts and aspect specific-images are captured. For the example ‘@ABQJournal @ABQJournalPhoto Bad accident at San Mateo and H751. Motorcycle hits car and flips

flip', the corresponding image is presented as Figure 3(a). The target and aspect in the sentence are 'San Mateo' and 'event', respectively. According to Figure 3(b), It can see that our model pays more attention to the motorcycle within the image. In addition, the MHA model (with Head = 4), assigns more attention weights to words like 'Motorcycle', 'bad' and 'accident', as shown in Figure 3(c). At this stage, our model classifies the sentiment polarity as negative, which demonstrates that our model can properly capture the information and interaction of multimodalities.

VI. CONCLUSION

In this work, the researchers present a novel multimodal sentiment analysis task, namely TABMSA. As such, in line with the TABMSA tasks, the EF-Net model is designed and deployed. The researchers first construct the representations of multimodal inputs. By employing the MHA network, the interaction between different representations is precisely captured to deliver more-related information. Moreover, the targeted aspect representation is enriched with the fusion of context and image information, which improves the multimodal sentiment classification accuracy to a large extent. Experiments results validate that the proposed model stably outperforms the baseline models.

ACKNOWLEDGMENT

(Donghong Gu and Jiaqian Wang contributed equally to this work.)

REFERENCES

- [1] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.
- [2] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 371–378.
- [3] Y. Yu, H. Lin, J. Meng, and Z. Zhao, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks," *Algorithms*, vol. 9, no. 2, p. 41, Jun. 2016.
- [4] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 929–932.
- [5] J. Yu and J. Jiang, "Adapting BERT for target-oriented multimodal sentiment classification," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5408–5414.
- [6] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 429–439, 2020.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [8] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel, "SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, 2016, pp. 1–11.
- [9] J. Huang, Y. Xue, X. Hu, H. Jin, X. Lu, and Z. Liu, "Sentiment analysis of Chinese online reviews using ensemble learning framework," *Cluster Comput.*, vol. 22, no. S2, pp. 3043–3058, Mar. 2019.
- [10] H. Li, Y. Xue, H. Zhao, X. Hu, and S. Peng, "Co-attention networks for aspect-level sentiment analysis," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, Dunhuang, China, 2019, pp. 1–10.
- [11] W. Haiming, Z. Yue, J. Xi, X. Yun, and W. Ziwen, "Shared-private LSTM for multi-domain text classification," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, Dunhuang, China, 2019, pp. 116–128.
- [12] H. Li, T. Yuan, H. Wu, Y. Xue, and X. Hu, "Granular computing-based multi-level interactive attention networks for targeted sentiment analysis," *Granular Comput.*, vol. 5, no. 3, pp. 387–395, 2020.
- [13] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 5876–5883.
- [14] Z. Ye and Z. Li, "A variant of recurrent entity networks for targeted aspect-based sentiment analysis," in *Proc. ECAI*, 2020, pp. 2268–2274.
- [15] F. Liu, T. Cohn, and T. Baldwin, "Recurrent entity networks with delayed memory update for targeted aspect-based sentiment analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 278–283.
- [16] H. Hong and J. Song, "Utilizing BERT for detecting aspect categories on TABSA via adjusting self-attention among words," in *Proc. Int. Conf. Intell. Comput. Hum.-Comput. Interact. (IHCICI)*, Dec. 2020, pp. 66–70.
- [17] S.-H. Wang, "AVNC: Attention-based VGG-style network for COVID-19 diagnosis by CBAM," *IEEE Sensors J.*, early access, Feb. 26, 2021, doi: 10.1109/JSEN.2021.3062442.
- [18] S.-H. Wang, Q. Zhou, M. Yang, and Y.-D. Zhang, "ADVIAN: Alzheimer's disease VGG-inspired attention network based on convolutional block attention module and multiple way data augmentation," *Frontiers Aging Neurosci.*, vol. 13, Jun. 2021, Art. no. 687456.
- [19] Y.-D. Zhang, Z. Dong, S. H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, and F. J. Martínez, "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation," *Inf. Fusion*, vol. 64, pp. 149–187, Dec. 2020.
- [20] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
- [21] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1–11.
- [22] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1–7.
- [23] F. Fan, Y. Feng, and D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3433–3442.



DONGHONG GU received the B.S. degree from South China Normal University, China, in 2019, where he is currently pursuing the M.S. degree with the School of Physics and Telecommunication Engineering. His research interests include nature language processing, text classification, and sentiment analysis.



JIAQIAN WANG received the M.S. degree from South China Normal University, Guangzhou, China, in 2020. He is currently working as a Research Assistant with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include artificial intelligence, nature language processing, photoacoustic holographic imaging, and physics informed deep neural networks.



SHAOHUA CAI graduated from Jinan University, in 2002. She is currently a Senior Experimenter with South China Normal University. Her research interests include data mining and natural language processing.



CHI YANG received the B.S. degree from Central South University for Nationalities, China, in 2019. She is currently pursuing the M.S. degree with the School of Physics and Telecommunication Engineering, South China Normal University. Her research interests include nature language processing, text classification, and sentiment analysis.



ZHENGXIN SONG received the B.S. degree from Wuhan Polytechnic University, China, in 2020. He is currently pursuing the M.S. degree with the School of Physics and Telecommunication Engineering, South China Normal University. His research interests include nature language processing, text classification, and sentiment analysis.



HAOLIANG ZHAO received the B.S. degree from Guangdong Medical University, China, in 2019. He is currently pursuing the M.S. degree with the School of Physics and Telecommunication Engineering, South China Normal University. His research interests include nature language processing, text classification, and sentiment analysis.



LUWEI XIAO (Member, IEEE) received the M.S. degree from South China Normal University, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, East China Normal University, China. His research interests include natural language process, machine learning, and sentiment analysis. He is a member of the CCF.



HUA WANG (Senior Member, IEEE) received the Ph.D. degree from the University of Southern Queensland, Australia. He was a Professor with the University of Southern Queensland before joined Victoria University. He has more than ten years teaching and working experience in applied informatics at both enterprise and university. He has expertise in electronic commerce, business process modeling, and enterprise architecture. He is currently a full time Professor with Victoria University. As a Chief Investigator, three Australian Research Council (ARC) Discovery grants have been awarded, since 2006, and 280 peer-reviewed scholar articles have been published. Ten Ph.D. students have already graduated under his principal supervision.

...