Metadata of the chapter that will be visualized in SpringerLink

Book Title	Artificial Intelligence	
Series Title		
Chapter Title	Interactive Fusion Network with Recurrent Attention for Multimodal Aspect-based Sentiment Analysis	
Copyright Year	2023	
Copyright HolderName	The Author(s), under exclusive license to Springer Nature Switzerland AG	
Author	Family Name	**Wang**
	Particle	
	Given Name	**Jun**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Harbin Institute of Technology (Shenzhen)
	Address	Shenzhen, China
	Division	
	Organization	Joint Lab of HIT-KONKA
	Address	Shenzhen, China
	Email	wjun.hit@gmail.com
	ORCID	http://orcid.org/0000-0002-7581-8996
Author	Family Name	**Wang**
	Particle	
	Given Name	**Qianlong**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Harbin Institute of Technology (Shenzhen)
	Address	Shenzhen, China
	Division	
	Organization	Joint Lab of HIT-KONKA
	Address	Shenzhen, China
	Email	qlwang15@outlook.com
	ORCID	http://orcid.org/0000-0002-3011-0580
Author	Family Name	**Wen**
	Particle	
	Given Name	**Zhiyuan**
	Prefix	
	Suffix	
	Role	
	Division	

	Organization	Harbin Institute of Technology (Shenzhen)
	Address	Shenzhen, China
	Division	
	Organization	Joint Lab of HIT-KONKA
	Address	Shenzhen, China
	Email	wenzhiyuan2012@gmail.com
	ORCID	http://orcid.org/0000-0003-4106-1312
Author	Family Name	**Liang**
	Particle	
	Given Name	**Xingwei**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Konka Research Institute
	Address	Shenzhen, China
	Division	
	Organization	Joint Lab of HIT-KONKA
	Address	Shenzhen, China
	Email	21BF51014@stu.hit.edu.cn
Corresponding Author	Family Name	**Xu**
	Particle	
	Given Name	**Ruifeng**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Harbin Institute of Technology (Shenzhen)
	Address	Shenzhen, China
	Division	
	Organization	Joint Lab of HIT-KONKA
	Address	Shenzhen, China
	Email	xuruifeng@hit.edu.cn
	ORCID	http://orcid.org/0000-0001-9885-2364

Abstract	The goal of multimodal aspect-based sentiment analysis is to comprehensively utilize data from different modalities (*e.g.,*, text and image) to identify aspect-specific sentiment polarity. Existing works have proposed many methods for fusing text and image information and achie-ved satisfactory results. However, they fail to filter noise in the image information and ignore the progressive learning process of sentiment features. To solve these problems, we propose an interactive fusion network with recurrent attention. Specifically, we first use two encoders to encode text and image data, respectively. Then we use the attention mechanism to obtain the semantic information of the image at the token level. Next, we employ GRU to filter out the noise in the image and fuse information from different modalities. Finally, we design a decoder with recurrent attention to progressively learn aspect-specific sentiment features for classification. The results on two Twitter datasets show that our method outperforms all baselines.

Keywords (separated by '-')	Multimodal aspect-based sentiment analysis - Attention mechanism - Progressively learning

Interactive Fusion Network with Recurrent Attention for Multimodal Aspect-based Sentiment Analysis

Jun Wang[1,3], Qianlong Wang[1,3], Zhiyuan Wen[1,3], Xingwei Liang[2,3], and Ruifeng Xu[1,3(✉)]

[1] Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2] Konka Research Institute, Shenzhen, China
[3] Joint Lab of HIT-KONKA, Shenzhen, China
21BF51014@stu.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract. The goal of multimodal aspect-based sentiment analysis is to comprehensively utilize data from different modalities (*e.g.,*, text and image) to identify aspect-specific sentiment polarity. Existing works have proposed many methods for fusing text and image information and achieved satisfactory results. However, they fail to filter noise in the image information and ignore the progressive learning process of sentiment features. To solve these problems, we propose an interactive fusion network with recurrent attention. Specifically, we first use two encoders to encode text and image data, respectively. Then we use the attention mechanism to obtain the semantic information of the image at the token level. Next, we employ GRU to filter out the noise in the image and fuse information from different modalities. Finally, we design a decoder with recurrent attention to progressively learn aspect-specific sentiment features for classification. The results on two Twitter datasets show that our method outperforms all baselines.

AQ1

Keywords: Multimodal aspect-based sentiment analysis · Attention mechanism · Progressively learning

1 Introduction

Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis task whose purpose is to identify the sentiment polarity corresponding to a particular aspect term. For example, in the text "*the dishes taste good, but the queue time is too long*", the sentiment polarities of "*dishes*" and "*queue time*" are positive and negative, respectively.

With the rapid development of mobile networks, more people use multimodal content (the most common form is image-text pair) on social platforms instead of just text. Analyzing the sentiment contained in multimodal data has become an important research direction. Multimodal aspect-based sentiment analysis (MABSA) comprehensively considers data from multiple modalities to judge

L. Fang et al. (Eds.): CICAI 2022, LNAI 13606, pp. 1–12, 2022.
https://doi.org/10.1007/978-3-031-20503-3_24

the sentiment polarity of a specific aspect. Taking Fig. 1 as an example, if we only consider the text, we cannot clearly judge the sentiment polarity of the aspect term "*Cuba*". However, if we take into account the beautiful scenery in the image, it is easy to conclude that the sentiment polarity of "*Cuba*" is positive.

Fig. 1. An example of MABSA. The aspect is marked in orange. (Color figure online)

To address the challenges on ABSA posed by multimodal data, many researchers have proposed solutions from different perspectives. For example, Xu et al. [18] proposed a multi-interactive memory network to capture the interaction between text and image. Besides, Zhang et al. [21] designed a discriminant matrix to fuse the information in images and texts. Zhou et al. [22] designed an adversarial training strategy to align the semantic space of image and text representations, resulting in better multimodal feature fusion results. Although the above methods have achieved satisfactory results, they still have the following shortcomings: (1) they directly interact text with image information without considering image noise. (2) they model MABSA as a classification task without considering the recurrent progressive learning of sentiment features.

To solve these two shortcomings, we propose an interactive fusion network with recurrent attention (IFNRA) for MABSA. Our model consists of three parts: encoder, interactive fusion module, and decoder with recurrent attention. To be specific, we first extract text and image features using BERT [3] and Bottom-Up Attention Model [1] (BUA), respectively. Then we use the attention mechanism to obtain the image semantic information of each token. The acquired image semantic information inevitably contains some noise. To filter out image noise and obtain multimodal representations, we use GRU [2] to fuse beneficial image information and textual information. Finally, we design a decoder with recurrent attention mechanism. Each cycle extracts required information from the multimodal representations to continuously update and improve the sentiment feature of the specific aspect. The classifier infers the sentiment polarity based on the output of the last cycle. We conduct experiments on the Twitter datasets. The results show that our method[1] can achieve better performance than baselines.

2 Related Work

Aspect-based Sentiment Analysis. Early research on ABSA mainly used traditional machine learning methods. Constructing features based on text and external knowledge such as lexical resources was the focus of works [7,9,10,16]. The selection and construction of features largely determined the performance of the model. In recent years, neural networks have achieved rapid development,

[1] The source code is publicly released at https://github.com/0wj0/IFNRA.

exhibiting powerful feature learning capabilities. Dong et al. [4] introduced recurrent neural networks to the ABSA and obtained sentiment polarities of aspect terms according to contextual and syntactic relations. Tang et al. [14] incorporated aspect information into LSTMs to establish connections between aspect words and context words without relying on external knowledge. Furthermore, Wang et al. [17] used the attention mechanism to obtain information that is highly correlated with the sentiment polarity of aspect words at the word and clause levels. With the development and widespread use of pre-trained language models, Phan et al. [12] used BERT for contextual embedding and combined part-of-speech information and syntactic information for ABSA.

Multimodal Aspect-based Sentiment Analysis. Different from traditional ABSA, MABSA comprehensively considers the impact of multi-modalities data on the sentiment polarity. The characteristics of social media data (such as the inconsistency between pictures and texts) pose certain challenges on MABSA. To solve the above problems, Xu et al. [18] first proposed the MABSA. They used the attention mechanism to extract information related to specific aspects in each modality, and designed a multi-hop memory network to make the information of each modality fully interact. Yu et al. [19] used additional BERT layers to obtain features of image regions that are closely related to aspects. In addition, Zhang et al. [21] designed a discriminant matrix to fuse the complementary information between different modalities, so that the representation can contain richer semantic information. Zhou et al. [22] designed an adversarial training strategy to align the semantic space of image and text representations, and then made any two of text, image and aspect interact with each other through the multimodal interaction layer.

Unlike previous works, we propose an interactive fusion network with recurrent attention. Our model filters noise in image through GRU and uses recurrent attention to progressively learn different sentiment features of specific aspects.

3 Model Architecture

3.1 Task Definition

Given a text $T = \{w_1, w_2, ..., w_n\}$, a related image I, and a specific aspect $A = \{a_1, a_2, ..., a_m\}$, the goal of the task is to predict the sentiment polarity $c \in C$ of the aspect. Here, A is a subsequence of T, n and m denote the number of tokens in T and A, respectively. $C = \{Negative, Neutral, Positive\}$.

3.2 Overview

As shown in Fig. 2, our proposed interactive fusion network consists of an encoder, an interactive fusion module, and a decoder with recurrent attention. We use BERT and BUA as encoders, which encode text and images into embedding representations, respectively. After obtaining representations of text and

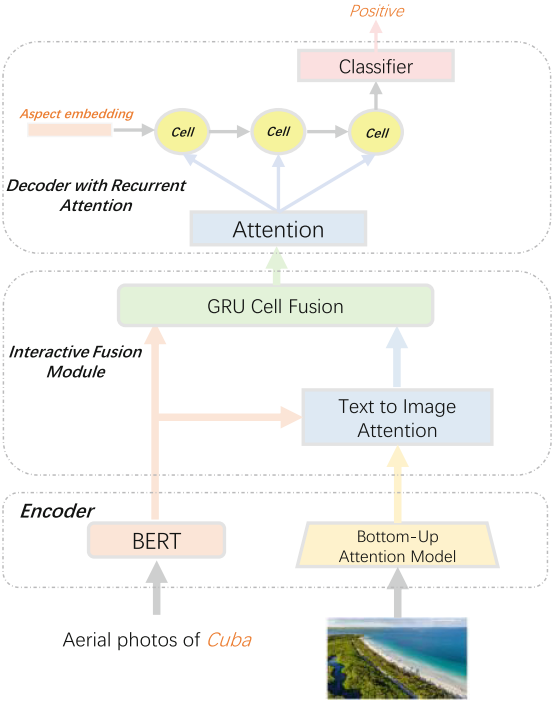


Fig. 2. The overview of our interactive fusion network.

images, we use an interactive fusion module to fuse features from different modalities to obtain multimodal representations. Finally, we complete the classification task using a decoder with the recurrent attention mechanism, which recurrently learns aspect-specific sentiment features and uses feature from the last step to judge the sentiment polarity.

3.3 Model Architecture

Encoder. For each token w_i in the text T, we use BERT to obtain its embedding $e_i^w \in \mathbb{R}^d$ containing contextual semantic information:

$$E^w = BERT(T) \tag{1}$$

The encoded text is $E^w = \{e_1^w, e_2^w, ..., e_n^w\}$, and the encoded aspect is $E^a = \{e_1^a, e_2^a, ..., e_m^a\}$. Here, E^a is a subsequence of E^w. d denotes the dimension of the hidden state of BERT.

The convolutional neural network evenly divides the image into several regions of equal size, and outputs the features of each region. However, this method destroys the semantic information expressed by the entire object or region in the original image. To solve the above problems, we choose the pre-trained BUA model as the image encoder. For the given image I, the BUA model

can detect objects and other salient regions in the image, and output the corresponding features. Then we use a linear layer to project the feature vectors into a d-dimensional space. The encoded image is $E^I = \{e_1^I, e_2^I, ..., e_k^I\}$. Here, $e_i^I \in \mathbb{R}^d$ denotes the feature vector of an object or region. k denotes the number of objects and regions in the image.

Interactive Fusion Module. It is challenging for MABSA to use the features of multiple modalities to obtain representations with richer and more complete semantic information. To obtain better multimodal representation, we let the features of different modalities interact and fuse at the token level. Specifically, we design a multimodal interactive attention mechanism. For each token vector e_i^w in the encoded text E^w, we use it to query the image features E^I to obtain relevant image information v_i^{img}.

$$v_i^{img} = Attention(e_i^w, E^I) \tag{2}$$

$$Attention(e_i^w, E^I) = softmax(\frac{Q_i^w (K^I)^T}{\sqrt{d}})V^I \tag{3}$$

$$Q_i^w, K^I, V^I = e_i^w \cdot W^{Qw}, E^I \cdot W^{KI}, E^I \cdot W^{VI} \tag{4}$$

Here, $W^{Qw} \in \mathbb{R}^{d\times d}$, $W^{KI} \in \mathbb{R}^{d\times d}$ and $W^{VI} \in \mathbb{R}^{d\times d}$ are learnable parameters. The obtained image information will inevitably contain noise. We use GRU to filter the noise and fuse multi-modal information. For token w_i, we take its token vector e_i^w and its image information v_i^{img} as the initial hidden state and input of the GRU cell, respectively. Its multimodal representation e_i^{wI} is the updated hidden state. The formula is as follows:

$$e_i^{wI} = GRUCell(v_i^{img}, e_i^w) \tag{5}$$

All token-level representations constitute the fusion result of multi-modal information, i.e., $E^{wI} = \{e_1^{wI}, e_2^{wI}, ..., e_n^{wI}\}$.

Decoder. To extract aspect-specific information from multimodal fusion representations, we design a decoder with recurrent attention, which considers the recurrent learning process of different attention features. Specifically, we take the average of all word vectors in the encoded aspect E^a as the initial aspect representation h_0. At the t-th time step, the multimodal fusion representation E^{wI} is queried with the aspect representation h_{t-1} produced at the previous time step to obtain the multimodal information v_t^{mul}. The aspect representation is continuously updated according to the acquired multimodal information through the GRU[2].

[2] We create an instance of GRU cell for each time step.

$$v_t^{mul} = Attention(h_{t-1}, E^{wI}), t \geq 1 \tag{6}$$

$$Attention(h_{t-1}, E^{wI}) = softmax(\frac{Q_{t-1}^h (K^{wI})^T}{\sqrt{d}})V^{wI}, t \geq 1 \tag{7}$$

$$Q_{t-1}^h = h_{t-1} \cdot W^{Qh}, t \geq 1 \tag{8}$$

$$K^{wI}, V^{wI} = E^{wI} \cdot W^{KwI}, E^{wI} \cdot W^{VwI} \tag{9}$$

$$h_t = GRUCell_t(v_t^{mul}, h_{t-1}), t \geq 1 \tag{10}$$

Here, $h_0 = \frac{1}{m}\sum_i e_i^a$, $W^{Qh} \in \mathbb{R}^{d \times d}$, $W^{KwI} \in \mathbb{R}^{d \times d}$ and $W^{VwI} \in \mathbb{R}^{d \times d}$ are learnable parameters. After updating and refining for N time steps, the multimodal aspect representation produced at the last time step is fed into a softmax layer to predict the probability distribution of its sentiment polarity.

$$p = softmax(W \cdot h_N + b) \tag{11}$$

Here, W and b are learnable parameters.

3.4 Loss Function and Optimizer

We take the cross-entropy between predicted results and targets as the loss function, which is calculated as follows:

$$loss = -\frac{1}{M}\sum_{j=1}^{M}\sum_{c \in C} q_j(c) log(p_j(c)) \tag{12}$$

Here, M denotes the total number of samples, $q_j(c)$ and $p_j(c)$ denote the true and predicted probability that sample j belongs to class c, respectively. We use the Adam [8] algorithm to minimize loss.

4 Experiments

4.1 Datasets and Settings

Datasets. We use the Twitter-15 and Twitter-17 datasets [19] to evaluate the effectiveness of the proposed method. The Twitter-15 and Twitter-17 are multimodal datasets where each sample contains a twitter text, a text-related image, a specific aspect, and its corresponding sentiment polarity label. There are three sentiment polarity labels: *positive*, *neutral*, and *negative*. The basic statistics of the Twitter-15 and Twitter-17 datasets are shown in Table 1.

Table 1. The basic statistics of datasets.

	Twitter-15			Twitter-17		
	Train	Dev	Test	Train	Dev	Test
Negative	368	149	113	416	144	168
Neutral	1883	670	607	1638	517	573
Positive	928	303	317	1508	515	493
Total	3179	1122	1037	3562	1176	1234
Avg. length	16.72	16.74	17.05	16.21	16.37	16.38

Experimental Settings. We select the pre-trained BERT-base[3] model as the text encoder. Its number of layers is 12, the size of the hidden layer state is 768, and the number of attention headers is 12. For the Bottom-Up Attention model, we use the pre-trained dynamic 10–100 model[4], setting the minimum and maximum number of features to 3 and 36, respectively. The BUA model is only used to extract image features and does not participate in training. The maximum value of time steps in the decoder is set to 3, *i.e.*, $N = 3$. Besides, the batch size is set to 32 and the learning rate is 2e-5. For each experiment, we run it 5 times with different random seeds and take the average as the final result. All experiments are performed on an NVIDIA GeForce RTX 3090 GPU. Following Yu et al. [19], we use accuracy and Macro-F1 as evaluation metrics.

4.2 Baselines

To verify the effectiveness of our model, we compare our model with several baseline methods:

(1) **RES-Target** uses ResNet [6] to encode images and BERT to encode aspect words, and then concatenates the encoding results of images and aspect words as features for sentiment polarity classification.
(2) **MGAN** [5] combines coarse- and fine-grained attention mechanisms to capture the interaction between aspect and context.
(3) **BERT** [3] uses a multi-layer transformer encoder [15] to obtain dynamic word vector representations.
(4) **BERT+BL** adds an additional BERT layer on top of the BERT-base model.
(5) **MIMN** [18] uses the attention mechanism to extract information in each modality, and designs a multi-hop memory network to make the information fully interact.
(6) **Res-MGAN** concatenates the image encoding results from ResNet and the text encoding results from MGAN.
(7) **Res-MGAN-TFN** uses tensor fusion network (TFN) [20] to fuse the image encoding results from ResNet and the text encoding results from MGAN.
(8) **Res-BERT+BL** concatenates the image encoding results from ResNet and the text encoding results from BERT+BL.

[3] https://huggingface.co/google/bert_uncased_L-12_H-768_A-12.
[4] https://github.com/MILVLG/bottom-up-attention.pytorch.

Table 2. Results of different models on the Twitter datasets. The best results are in **bold** and the second best results are *underlined*.

Modality	Model	Twitter-15		Twitter-17	
		Acc	Mac-F1	Acc	Mac-F1
Visual	RES-Target [6]	59.88	46.48	58.59	53.98
Text	MGAN [5]	71.17	64.21	64.75	64.16
	BERT [3]	74.15	68.86	68.15	65.23
	BERT+BL	74.25	<u>70.04</u>	68.88	66.12
Text+Visual	MIMN [18]	73.53	66.49	67.22	63.85
	Res-MGAN [5]	71.65	63.88	66.37	63.04
	Res-MGAN-TFN [20]	70.30	64.14	64.10	59.13
	Res-BERT+BL	<u>75.02</u>	69.21	<u>69.20</u>	<u>66.48</u>
	IFNRA(ours)	**76.03**	**71.79**	**70.78**	**69.48**

4.3 Main Results

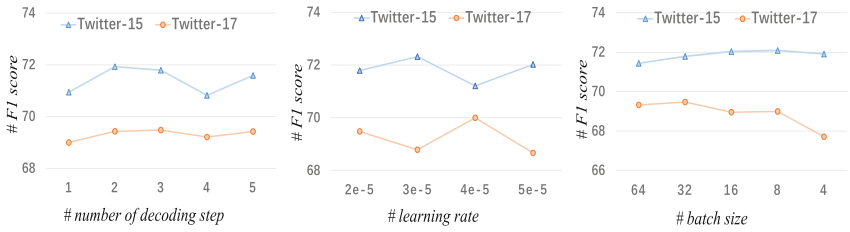
Table 2 presents performance comparison among different models. As can be seen, our model outperforms all baselines, which shows the effectiveness of the proposed model. Besides, we can find that the multimodal model achieves better accuracy than the single-modal model, which indicates the complementarity between the different modal data. Furthermore, the text-only models perform better than the image-only model, suggesting that MABSA has a greater dependence on text.

Table 3. Results of ablation experiments. The best results are in **bold** and the second best results are *underlined*.

Model	Twitter-15		Twitter-17	
	Acc	Mac-F1	Acc	Mac-F1
IFNRA	**76.03**	**71.79**	**70.78**	**69.48**
IFNRA *w/ txt2img_attn*	<u>75.60</u>	70.63	69.9	68.41
IFNRA *w/ fuse*	74.95	<u>70.65</u>	69.81	68.46
IFNRA *w/ classify*	74.97	70.61	<u>70.10</u>	<u>68.57</u>

4.4 Ablation Study

To verify the effectiveness of different modules, we conduct ablation experiments by constructing several variants: (1) **IFNRA** *w/ txt2img_attn* does not use GRU to fuse text and image information. It directly inputs the result $R = \{v_1^{img}, v_2^{img}, ..., v_n^{img}\}$ of Text-to-Image Attention to the decoder. (2) **IFNRA** *w/ fuse* directly adds the word vector e_i^w and the image information v_i^{img} to obtain the multimodal representation e_i^{wI} instead of going through the GRU.



(a) (b) (c)

Fig. 3. Hyper-parameters sensitivity analysis.

(3) **IFNRA** *w/ classify* directly uses the multimodal information v_1^{img} to judge the sentiment polarity without using recurrent attention.

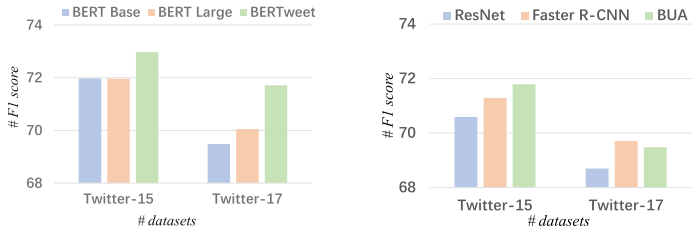
The results of ablation experiment are shown in Table 3. We find that the performance of the complete model is better than that of all variants, which shows that each module in IFNRA plays an important role. Specifically, comparing IFNRA *w/ txt2img_attn* with IFNRA can reflect the importance of interactive fusion of multimodal information. In addition, comparing IFNRA *w/ fuse* with IFNRA highlights the role of GRU in filtering noise and fusing multimodal data. Moreover, the effectiveness of the recurrent attention learning in decoder can be shown by comparing IFNRA *w/ classify* with IFNRA.

4.5 Discussion

Effect of Different Decoding Steps on Performance. To explore the effect of the number of decoding steps N on performance, we change N from 1 to 5. The experimental results are shown in Fig. 3a. We can find that the performance of the model gradually improves as the increase of the number of steps, which shows that increasing the number of steps in the recurrent decoding can obtain more effective attention features. Furthermore, when the number of steps continues to increase, the model performance peaks and then declines, indicating that increasing the number of steps too much can lead to too many parameters and over-fitting.

Effect of Different Learning Rates on Performance. To explore the effect of the learning rate on performance, we change the learning rate from 2e-5 to 5e-5. The experimental results are shown in Fig. 3b. As the learning rate changes, the model performance fluctuates only slightly. It means that a small change in the learning rate will not seriously affect the performance.

Effect of Different Batch Sizes on Performance. To explore the effect of the batch size on performance, we change the batch size from 4 to 64. The experimental results are shown in Fig. 3c. It can be found that when the batch



(a) Pre-trained language models. (b) Pre-trained vision models.

Fig. 4. The effect of pre-trained models on F1.

size is moderate (i.e. 32, 16 and 8), the overall performance is better. It indicates that a larger or smaller batch size could affect the generalizability of the model and the convergence of the parameters.

Effect of Different Pretrained Models on Performance. To explore the impact of different pre-trained models on performance, we replace different text and image encoders and conduct experiments. The results are shown in Fig. 4a and Fig. 4b. We can find that compared with BERT-base, although BERT Large[5] has a huge amount of parameters, the improvement brought by it is limited. We also note that BERTweet[6] [11] improves performance significantly, illustrating the importance of incorporating domain expertise into pre-trained language models. Furthermore, using Faster R-CNN[7] [13] as the image encoder can achieve better performance compared to using ResNet[8]. This shows that evenly dividing the image and extracting features will cause more loss of semantic information, and processing the objects or salient regions in the image as a whole can achieve better results. While BUA detects an object, it also predicts related properties (*e.g.*, color), which makes the image features contain more information. Therefore, BUA can achieve better overall performance than Faster R-CNN.

5 Conclusion

In this paper, we propose an interactive fusion network with recurrent attention to solve the MABSA. Specifically, we use BERT and BUA model to obtain features for the text and image, respectively. Furthermore, we design an interactive fusion module to obtain the semantic information of the image at the token level. We use GRU to filter out noise in image information and fuse data from different

[5] https://huggingface.co/bert-large-uncased.
[6] https://huggingface.co/vinai/bertweet-base.
[7] https://github.com/facebookresearch/detectron2/blob/main/configs/COCO-Detection/faster_rcnn_R_101_FPN_3x.yaml.
[8] https://download.pytorch.org/models/resnet50-0676ba61.pth.

modalities. Finally, we design a recurrent attention mechanism to progressively learn aspect-specific sentiment features from the fused multimodal representations. The feature obtained in the last step is used to predict the sentiment polarity. Our model outperforms all baselines on two Twitter datasets.

Acknowledgments.. This work was partially supported by the National Natural Science Foundation of China (61876053, 62006062, 62176076), Shenzhen Foundational Research Funding (JCYJ20200109113441941 and JCYJ2021032411 5614039), Joint Lab of HIT and KONKA.

References

1. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of CVPR, pp. 6077–6086 (2018). https://openaccess.thecvf.com/content_cvpr_2018/papers/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.pdf
2. Cho, K., van Merriënboer, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of EMNLP, pp. 1724–1734 (2014). https://aclanthology.org/D14-1179.pdf
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL, pp. 4171–4186 (2019). https://aclanthology.org/N19-1423/
4. Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., Xu, K.: Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of ACL, pp. 49–54 (2014). https://aclanthology.org/P14-2009.pdf
5. Fan, F., Feng, Y., Zhao, D.: Multi-grained attention network for aspect-level sentiment classification. In: Proceedings of EMNLP, pp. 3433–3442 (2018). https://aclanthology.org/D18-1380/?ref=githubhelp.com
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of CVPR, pp. 770–778 (2016). https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf
7. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: Proceedings of ACL, pp. 151–160 (2011). https://aclanthology.org/P11-1016.pdf
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of ICLR (Poster) (2015). https://openreview.net/forum?id=8gmWwjFyLj
9. Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S.: NRC-Canada-2014: detecting aspects and sentiment in customer reviews. In: Proceedings of SemEval, pp. 437–442 (2014). https://aclanthology.org/S14-2076.pdf
10. Mohammad, S., Kiritchenko, S., Zhu, X.: NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of SemEval, pp. 321–327 (2013). https://aclanthology.org/S13-2053.pdf
11. Nguyen, D.Q., Vu, T., Nguyen, A.T.: BERTweet: a pre-trained language model for English Tweets. In: Proceedings of EMNLP, pp. 9–14 (2020). https://aclanthology.org/2020.emnlp-demos.2.pdf
12. Phan, M.H., Ogunbona, P.O.: Modelling context and syntactical features for aspect-based sentiment analysis. In: Proceedings of ACL, pp. 3211–3220 (2020). https://aclanthology.org/2020.acl-main.293/?ref=githubhelp.com

13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS 28 (2015)
14. Tang, D., Qin, B., Feng, X., Liu, T.: Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING, pp. 3298–3307 (2016). https://aclanthology.org/C16-1311/?ref=githubhelp.com
15. Vaswani, A., et al.: Attention is all you need. In: NIPS 30 (2017). https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
16. Wagner, J., et al.: DCU: aspect-based polarity classification for Semeval task 4. In: Proceedings of COLING, pp. 223–229 (2014). https://aclanthology.org/S14-2.pdf#page=243
17. Wang, J., et al.: Aspect sentiment classification with both word-level and clause-level attention networks. In: Proceedings of IJCAI, vol. 2018, pp. 4439–4445 (2018). www.ijcai.org/proceedings/2018/0617.pdf
18. Xu, N., Mao, W., Chen, G.: Multi-interactive memory network for aspect based multimodal sentiment analysis. In: Proceedings of AAAI, vol. 33, pp. 371–378 (2019). https://ojs.aaai.org/index.php/AAAI/article/view/3807/3685
19. Yu, J., Jiang, J.: Adapting BERT for target-oriented multimodal sentiment classification. In: Proceedings of IJCAI (2015). www.ijcai.org/Proceedings/2019/0751.pdf
20. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of EMNLP, pp. 1103–1114 (2017)
21. Zhang, Z., Wang, Z., Li, X., Liu, N., Guo, B., Yu, Z.: ModalNet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. World Wide Web **24**(6), 1957–1974 (2021). https://link.springer.com/article/10.1007/s11280-021-00955-7
22. Zhou, J., Zhao, J., Huang, J.X., Hu, Q.V., He, L.: MASAD: a large-scale dataset for multimodal aspect-based sentiment analysis. Neurocomputing **455**, 47–58 (2021). www.sciencedirect.com/science/article/pii/S0925231221007931

Author Queries

Chapter 24

Query Refs.	Details Required	Author's response
AQ1	As per Springer style, both city and country names must be present in the affiliations. Accordingly, we have inserted the city and country names "Shenzhen, China" in the affiliation 3. Please check and confirm if the inserted city and country names are correct. If not, please provide us with the correct city and country names.	