



---

*Research article*

## **Self-adaptive attention fusion for multimodal aspect-based sentiment analysis**

**Ziyue Wang<sup>1,2</sup> and Junjun Guo<sup>1,2,\*</sup>**

<sup>1</sup> Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

<sup>2</sup> Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China

\* **Correspondence:** Email: [guojjgb@163.com](mailto:guojjgb@163.com).

**Abstract:** Multimodal aspect term extraction (MATE) and multimodal aspect-oriented sentiment classification (MASC) are two crucial subtasks in multimodal sentiment analysis. The use of pretrained generative models has attracted increasing attention in aspect-based sentiment analysis (ABSA). However, the inherent semantic gap between textual and visual modalities poses a challenge in transferring text-based generative pretraining models to image-text multimodal sentiment analysis tasks. To tackle this issue, this paper proposes a self-adaptive cross-modal attention fusion architecture for joint multimodal aspect-based sentiment analysis (JMABSA), which is a generative model based on an image-text selective fusion mechanism that aims to bridge the semantic gap between text and image representations and adaptively transfer a textual-based pretraining model to the multimodal JMASA task. We conducted extensive experiments on two benchmark datasets, and the experimental results show that our model significantly outperforms other state of the art approaches by a significant margin.

**Keywords:** natural language processing; sentiment analysis; joint multimodal aspect-based sentiment analysis; multimodal fusion; self-adaptive fusion

---

### **1. Introduction**

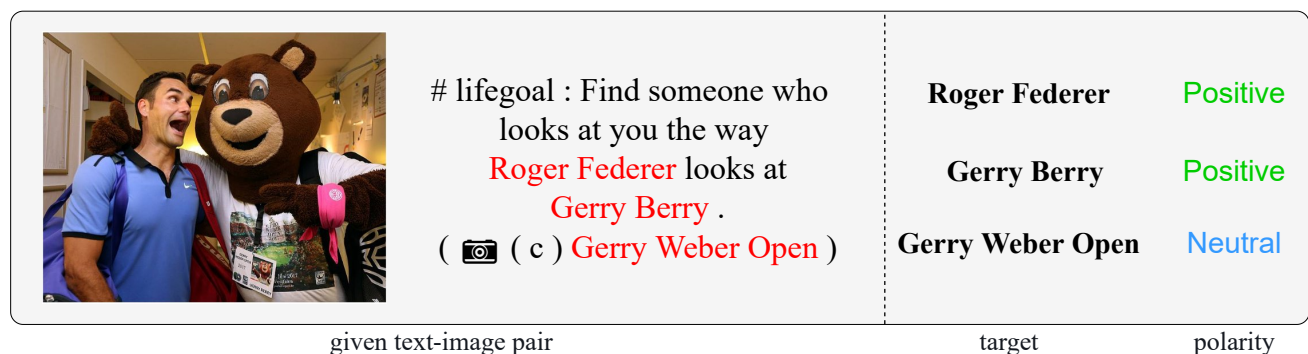
Early sentiment analysis mainly focused on text, considering only the interrelationships between words and phrases to analyze sentiment [1]. In recent years, with the content on internet social platforms gradually shifting from purely text-based content to multimodal content, the task of multimodal sentiment analysis has received increasing attention. Common multimodal sentiment analysis tasks include video sentiment analysis [2] and image-text sentiment analysis.

As a fundamental sentiment analysis task, joint multimodal aspect-based sentiment analysis

(JMABSA) aims to extract the potential aspect terms and identify aspects' sentiment polarities simultaneously from text with the aid of images, which has received increasing attention over the past few years. For example, in Figure 1, the objective of JMABSA is to detect all aspect-sentiment pairs, i.e., (Roger Federer, Positive), (Gerry Berry, Positive) and (Gerry Weber Open, Neutral).

Multimodal aspect term extraction (MATE) and multimodal aspect sentiment classification (MASC) are two types of subtasks contained in JMABSA. Most previous works prefer to cast JMABSA as the aforementioned two pipeline sub-tasks. However, this kind of step-by-step operation requires propagate artifacts generated in the first step to the next step, thereby reducing the sentiment analysis performance of the final results.

There have been many attempts to explore MATE and MASC based on pretrained models. Yu et al. [3] proposed a multimodal bidirectional encoder representation from transformers (BERT) for target-oriented sentiment classification by capturing the multimodal interactions with a target attention mechanism. Khan et al. [4] introduced a two-stream multimodal target sentiment classification model with BERT by combining text and image captions. Yu et al. [5] designed a hierarchical interactive multimodal transformer to identify the aspect-oriented sentiment polarities by capturing text-image interactions. Ling et al. [6] presented a task-specific vision-language pretrained model based on bidirectional and auto-regressive transformers (BART) [36] for MASC. Unfortunately, there are considerable feature representation gaps, as visual and textual features are initialized with their corresponding modality-specific models. Therefore, it will inevitably suffer from modality alignment ambiguity by directly incorporating visual features into the pretrained textual models.



**Figure 1.** An example of the MABSA task.

Image is another kind of modality data that contains many helpful details, such as the salient objects, the scenario information, the facial expression, etc. These visual details are valuable for aspect extraction and sentiment polarity identification, as depicted in Figure 1, it is challenging to determine the sentiment polarity of aspects based on only text information. However, the visual modality offers valuable clues, such as facial expressions, that assist in predicting the sentiment of Roger Federer and Gerry Berry. More concretely, in the MATE task, we prefer to capture the salient objects and the scenario information to enhance the aspect term extraction performance. In the MASC task, the facial expression is much helpful in identifying semantic polarities. Therefore, visual information could be employed as the pivotal information to bridge the task gap between MATE and MASC, eliminating the error propagation problem of JMABSA.

Although there is a significant modality gap between image and text, they can complement each other. Inspired by this phenomenon, this paper develops a visual-textual interactive sequence to se-

quence (Seq2Seq) framework based on BART to address the joint aspect term extraction and aspect-oriented sentiment classification problems. The inclusion of image information is indispensable for the JMABSA task. However, the semantic gap between text and image poses a challenge, impeding the effective integration of the two modalities in the multimodal BART model. Addressing the modality gap between text and image and establishing connections between the two modalities is of paramount importance for the JMABSA task. To address the semantic gap problem, this paper proposes an adaptive visual-to-textual fusion module to bridge the modality gap. The contributions of this work are summarized as follows.

- To eliminate the inherent semantic gap between textual and visual modalities, we employ image as pivotal information to bridge the semantic gap between textual and visual modalities, and visual details are dynamically extracted to enhance the performance of JMABSA.
- An adaptive visual-to-textual fusion module is built to adaptively incorporate task-specific visual information into a pretrained BART encoder to promote the network to learn a multimodal representation.
- Experiment results on TWITTER-15 and TWITTER-17 datasets show that the proposed approach significantly enhances the performance of MATE and MASC and improves F1 scores on two test sets. Moreover, our model almost achieves the performance of task-specific pretrained methods.

## 2. Related work

### 2.1. Text-based aspect based sentiment analysis (ABSA)

Aspect based sentiment analysis (ABSA) aims to identify sentiment polarities at the aspect level. In order to handle ABSA in different scenarios, there exists several subtasks in ABSA. The main research line of ABSA focuses on two primary subtasks: Aspect term extraction, and aspect sentiment classification. For aspect term extraction, some early works mainly focus on extracting sequence features via sequence tagging methods based on convolutional neural networks (CNN) [7] and recurrent neural networks (RNN) [8]. Recent works have discussed Seq2Seq methods on aspect term extraction [9, 10]. Similarly, for aspect sentiment classification, early studies were mainly based on manually designed features [11, 12]. In recent years, various deep learning approaches have been proposed, including attention-based methods [13–17], CNN-based networks [18, 19] and graph neural networks (GNN) based methods [20–24]. Concurrently, the pretrained language model BERT [25] has demonstrated exceptional performance across numerous natural language processing (NLP) tasks. Li et al. [26] achieved favorable results by employing the BERT model for aspect-based sentiment classification. Since these two subtasks are highly dependent on each other, more recent studies attempt to solve these two subtasks jointly.

Joint aspect sentiment analysis (JASA) aims to extract aspect and predict their sentiments jointly. Some studies leveraged the pipeline method to solve this problem [27, 28], which formulates the target extraction task as a sequence tagging problem. Hu et al. [29] proposed a span-based extract-then-classify framework. Recently, Yan et al. [30] proposed a unified generative framework based on BART, and achieved the state of the art performance on JASA. Despite achieving remarkable improvement, all the above studies only focus on the textual modality but fail to model the visual guidance for both subtasks. In our work, we aim to propose a multimodal architecture to handle both subtasks jointly.

## 2.2. Multimodal aspect-based sentiment analysis (MABSA)

In the past few years, MABSA has drawn much attention. Existing studies on MABSA mostly focus on the two subtasks of MABSA: MATE and MASC. As a pioneer, Xu et al. [31] first proposed the task of MASC. Several studies have focused on modeling the interactions among the aspect, text and image based on attention mechanisms [31–34]. With the successful application to tasks in NLP, Yu et al. [3] proposed a multimodal BERT architecture [25], which adapts BERT to obtain textual features and interactions among textual and visual modalities. Moreover, Khan et al. [4] adapted a transformer architecture for image caption, which translates the image input to an auxiliary sentence, then feeds the auxiliary sentence into a BERT language model. Despite these advances of methods in MABSA, almost all of them focus on handling each subtask independently, which ignores the innate connection between these two subtasks. Therefore, we aim to extend this line of research by proposing a more effective method that jointly performs MATE and MASC.

In recent years, inspired by the success of the JASA tasks, Ju et al. [35] introduced the task of joint multimodal aspect-sentiment analysis, which aims to jointly extract aspect and predict their sentiments from a text-image pair. More recently, Ling et al. [6] proposed a task-specific vision-language pretraining (VLP) framework for MABSA, which is a unified multimodal encoder-decoder architecture based on BART. Nevertheless, VLP failed to capture the alignment of between text and image modalities while transferring textual based generative pretraining models to image-text multimodal sentiment analysis task. In contrast to VLP, our proposed model aims to bridge the semantic gap between text and image representations and transfer textual-based pretraining models to the JMABSA task self-adaptively.

## 3. Methodology

Our proposed self-adaptive attention fusion (SAAF) model mainly focuses on bridging an effective modal to bridge the semantic gap between text and image. As shown in Figure 2, The SAAF comprises of several parts: Feature extraction, adaptive visual-to-textual fusion layer, and visual-enhanced BART module.

**Task definition.** We conceptualize the JMABSA task as a sequence labeling problem. Consider  $D$  as a set comprising multimodal samples. Formally, we are given a multimodal tweet comprising an image denoted as  $V$  and a sentence with  $n$  words denoted as  $T = (t_1, t_2, \dots, t_n)$ . Our goal is to obtain the sequence  $y$  that represents all potential aspect terms along with their respective sentiment polarities. We formulate the output as  $y = (a_1^b, a_1^e, s_1, \dots, a_i^b, a_i^e, s_i, \dots, a_k^b, a_k^e, s_k)$ , where  $a_i^b$  and  $a_i^e$  denote the beginning index and the end index of the  $i$ -th aspect,  $s_k$  denotes the sentiment polarity toward the aspect and  $k$  represents the number of aspect terms contained in  $T$ .

### 3.1. Feature extraction

**Text embedding.** Given the competitive performance exhibited by the Seq2Seq pretrained model BART [36] in the context of JASA [30], its utilization is adopted for acquiring word embeddings. In adherence to the procedure delineated in [6], the markers  $\langle s \rangle$  and  $\langle /s \rangle$  are employed to denote the initiation and termination of a sentence. Formally, the textual representation of a sample is denoted as:

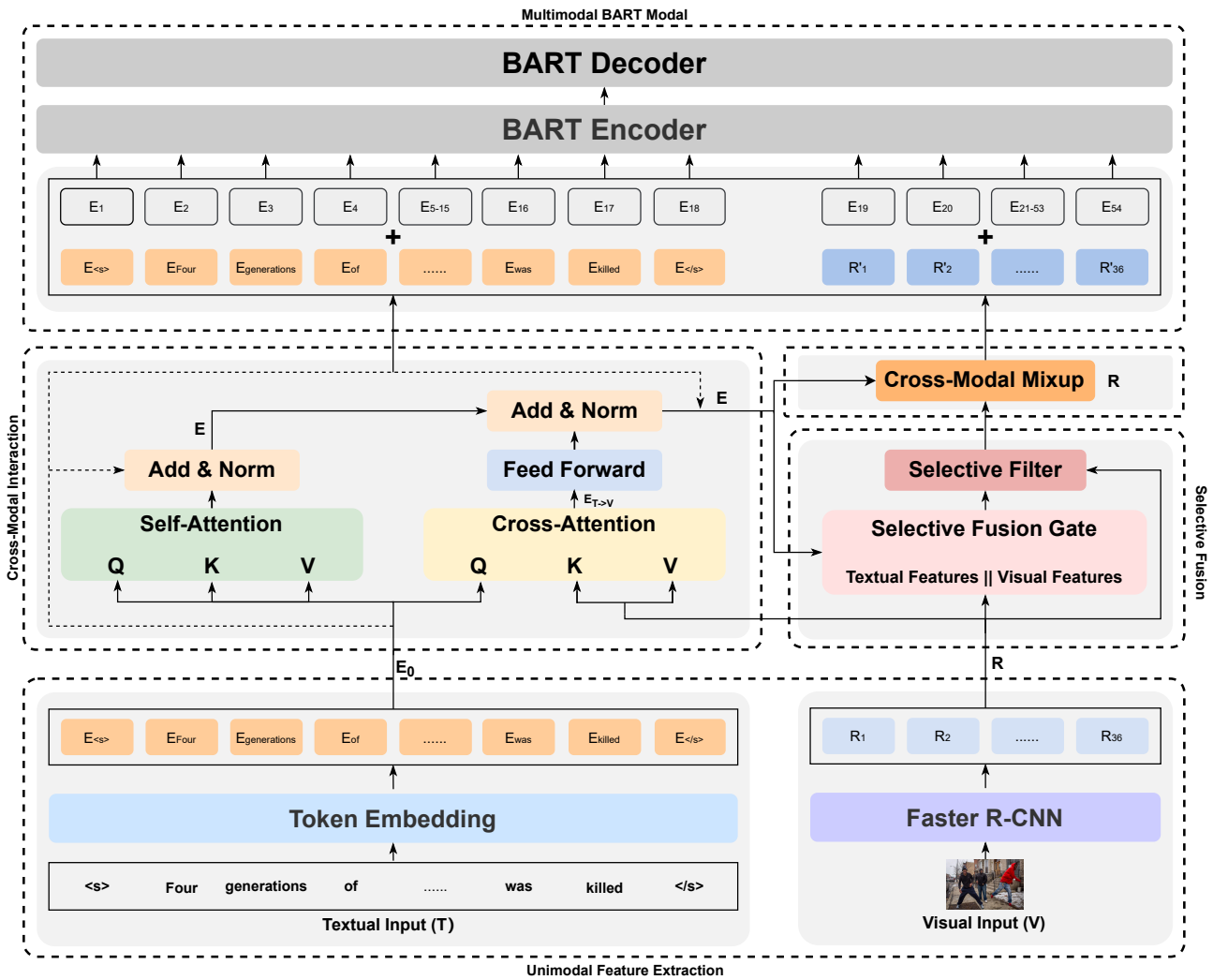


Figure 2. Overview of our SAAF model.

$$E_0 \in \mathbb{R}^{T \times d}, \tag{3.1}$$

where  $d$  denotes the dimension of BART, which is equal to 768.

**Image embedding.** The regional representations are obtained by Faster R-CNN [37]. Precisely, Faster R-CNN is adopted to extract all object proposals from an image denoted as  $V$ . Subsequently, 36 object proposals with the utmost confidence are retained. The identified object and its corresponding semantic significance are denoted as follows:

$$R = \text{FasterR} - \text{CNN}(V), \tag{3.2}$$

where *FasterR – CNN* denotes the Faster R-CNN [37] and  $R$  denotes the visual features:

$$R \in \mathbb{R}^{36 \times 2048}, \tag{3.3}$$

Then, the visual features are projected to match the textual embedding size of BART. Finally, the visual sequence is denoted as follows:

$$R \in \mathbb{R}^{36 \times d}. \quad (3.4)$$

### 3.2. Adaptive visual-to-textual fusion layer

**Cross-modal interaction.** The multi-head self-attention layer [38] is utilized to capture intra-modal interactions within the text. This is achieved by aggregating information from nearby words through text self-attention:

$$E = \text{Norm}(E_0 + \text{ATT}_{\text{self}}(E_0)), \quad (3.5)$$

where  $\text{ATT}_{\text{self}}$  denotes the multi-head self-attention, the textual feature is set as the query/key/value matrix and  $\text{Norm}$  denotes the layer normalization [39].

Simultaneously, a cross-modal transformer layer [40] is utilized to achieve inter-modal interaction across text and visual modalities. In this context, the textual features  $E$  function as the query matrix, while the visual features  $R$  serve as the key/value matrix, resulting in the following relationship:

$$E_{T \rightarrow V} = \text{ATT}_{\text{cross}}(E_0, R), \quad (3.6)$$

where  $\text{ATT}_{\text{cross}}$  denotes the cross-modal transformer.

Subsequently,  $E_{X \rightarrow V}$  is input into a feed-forward network (FFN) followed by a normalization layer. To enhance the textual representation further, an additional residual connection is established from  $E$ :

$$E = \text{Norm}(E + \text{FFN}(E_{X \rightarrow V})), \quad (3.7)$$

where  $\text{FFN}$  denotes a feed-forward network [38].

Visual information and textual information are merged through cross-modal interaction. Compared with previous work, our proposed approach can better extract text closely related visual features better in JMABSA.

**Selective fusion.** With the strengthened textual representation obtained through cross-modal interaction, the selective fusion further aims to filter out unrelated region features for the text. Essentially, the selective fusion receives two inputs: One is the strengthened textual representation  $E$ , and the other is purely visual feature  $R$ . Initially, a concatenation of  $R$  and  $E$  is performed to generate a bimodal factor denoted as  $[R; E]$ . This factor is then employed to compute the gate vector  $g$ :

$$g = \text{sigmoid}(\text{Linear}([R; E])) \quad (3.8)$$

where  $\text{sigmoid}$  denotes a Sigmoid nonlinear activation function.

The selective fusion gate highlights the relevant information within the visual modality conditioned on the textual representation that encompasses image information. Subsequently, the gate vector is utilized to acquire the textually related regional feature  $R$  through the application of the selective filter:

$$R = g * R. \quad (3.9)$$

**Cross-modal mixup.** To enhance the resilience of multimodal representation, the cross-modal mixup model is devised. The core philosophy behind cross-modal mixup is to create new samples by linearly interpolating a pair of training samples to exhibit linearity within the training data. A particularly appealing implementation of such multimodal data augmentation approach is studied in TMix [41]. The synthetic sample is generated as follows:

$$R = \lambda R + (1 - \lambda)E, \quad (3.10)$$

where  $\lambda$  is a scalar of balancing textual features and visual features, sampled from a Beta ( $\alpha, \beta$ ) distribution:

$$\lambda \sim Be(\alpha, \beta), \quad (3.11)$$

where  $Be$  denotes the Beta distribution and  $\alpha$  and  $\beta$  denote the hyperparameter to control the distribution of  $\lambda$ .  $R$  is produced as the ultimate visual representation.

### 3.3. Visual-enhanced BART module

The backbone of our model is BART [36], which is a Transformer-based autoencoder for Seq2Seq model. Following [6, 42], the original BART model is transformed into a multimodal variant capable of encoding the multimodal input.

**Encoder.** The encoder of our model is based on a multilayer bidirectional Transformer. Following [42], two distinct tokens,  $\langle \text{img} \rangle$  and  $\langle / \text{img} \rangle$ , are introduced to signify the initiation and culmination of visual features generated by the multimodal interpolation layer. Subsequently, we postulate that the original text representation  $E$  and the visual representation enriched with multimodal information  $R$  constitute the multimodal output denoted as  $D$ :

$$D = E_0 \oplus R, \quad (3.12)$$

where  $\oplus$  denotes the concatenation operation.

Following this,  $D$  is input into the position embedding layer to derive the ultimate multimodal representation:

$$D = \text{Dropout}(\text{Norm}(PE(D) + D)), \quad (3.13)$$

where  $D \in R^{(T+36) \times d}$  and  $PE$  denotes the position embedding layer.

Finally,  $D$  serves as the input for the multimodal BART encoder.

**Decoder.** The decoder of our model is also a multilayer Transformer. Different from the bidirectional encoder, the decoder is unidirectional. The output of the multimodal BART encoder is denoted as  $H_m$ .

$$H_m = \text{Encoder}(D) \quad (3.14)$$

The predict distribution as follows:

$$P(\theta) = \text{Softmax}(MLP(H_m)), \quad (3.15)$$

where *MLP* denotes the multilayer perceptron.

The loss function is determined by calculating the cross-entropy between the predicted label distribution and the true label distribution during the training process:

$$\mathcal{L} = -\mathbb{E}_{X \sim D} \log P(\theta|X), \quad (3.16)$$

where  $\theta$  denotes the true sentiment provided in the dataset and  $X$  denotes the multimodal input.

## 4. Experiments

### 4.1. Dataset

Two benchmark datasets, TWITTER-15 and TWITTER-17, are employed for evaluation as per the reference [3]. The detailed statistics of both datasets are shown in Table 1.

**Table 1.** Statistics of two benchmark datasets for JMABSA.

|               | TWITTER-15 |      |      | TWITTER-17 |      |      |
|---------------|------------|------|------|------------|------|------|
|               | Train      | Dev  | Test | Train      | Dev  | Test |
| Positive      | 928        | 303  | 317  | 1508       | 515  | 493  |
| Neutral       | 1883       | 670  | 607  | 1638       | 517  | 573  |
| Negative      | 368        | 149  | 113  | 416        | 144  | 168  |
| Total Aspects | 3179       | 1122 | 1037 | 3562       | 1176 | 1234 |
| Sentence      | 2101       | 727  | 674  | 1746       | 577  | 587  |

### 4.2. Evaluation metrics

The evaluation metrics employed to assess the performance include micro F1 score (F1), precision (P), and recall (R). The micro F1 score combines the precision and recall of the model, providing a comprehensive assessment of the overall performance. Precision measures the model's ability to correctly predict positive class samples, while recall gauges the model's success in capturing positive class samples. The integrated use of these three metrics contributes to a thorough evaluation of the model's performance in multi-class classification tasks, offering insights into different aspects of its effectiveness.

### 4.3. Implementation details

Our approach is implemented using PyTorch (version torch-1.11.0) on hardware comprising an RTX 3070Ti. The hidden size of our model is 768, which is the same as the dimension in BART [36]. The training of the model is conducted with the implementation of the early stopping mechanism to prevent overfitting. In particular, the training process spans 100 epochs, during which the model's performance on the validation set is assessed at each epoch. The training ceases if the model fails to exhibit improved F1 scores on the validation set for  $p$  consecutive epochs, where  $p$  is a predefined hyperparameter. Subsequently, the final model is derived from the last checkpoint, and its performance is evaluated using the test set.



#### 4.4. Baselines

Our primary focus revolves around comparing our SAAF model against two distinct categories of existing baseline systems using our proposed methodology.

Our analysis first commences with the evaluation of text-only methodologies: 1) SPAN [29] is a span-based method that formulates the JASA task as a span prediction problem, 2) directional graph convolutional networks (D-GCN) [43] proposes a BERT-based graph convolution network that formulates the JASA task as a sequence labeling problem to leverage synaptic information between words and 3) BART [30] is a unified generative framework based on BART that formulates the JASA task as an index generation problem.

Additionally, the following multimodal strategies are taken into account for JMASA since there are few studies for JMASA. 1) Initially, two pipeline approaches are executed using representative methods of MATE and MASC: unified multimodal transformer (UMT)+TomBERT and OSCGA+TomBERT, 2) UMT-collapsed [44], OSCGA-collapsed [45] and relation propagation-based BERT (RpBERT)-collapsed [46] are three collapsed tagging approaches, 3) JML [35] is the first multimodal joint learning approach, which proposed an auxiliary relation detection module to control the exploitation of visual information, 4) VLP-MABSA [6], which is a unified multimodal encoder-decoder architecture for multimodal joint learning method and 5) cross-modal multitask transformer (CMMT) [47], which proposed a text-guided cross-modal interaction module to dynamically control the contributions of visual information.

**Table 2.** Comparison between previous methods and our SAAF model on two benchmark datasets. <sup>a</sup> denotes the results from Ju et al. <sup>b</sup> denotes the results are from Liang et al. <sup>c</sup> denotes the results from Yang et al.

|                               | TWITTER-15  |             |             | TWITTER-17  |             |             |
|-------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                               | P           | R           | F1          | P           | R           | F1          |
| Text-based methods            |             |             |             |             |             |             |
| SPAN <sup>a</sup>             | 53.7        | 53.9        | 53.8        | 59.6        | 61.7        | 60.6        |
| D-GCN <sup>a</sup>            | 58.3        | 58.8        | 59.4        | 64.2        | 64.1        | 64.1        |
| BART <sup>b</sup>             | 62.9        | 65.0        | 63.9        | 65.2        | 65.6        | 65.4        |
| Multimodal methods            |             |             |             |             |             |             |
| UMT+TomBERT <sup>a</sup>      | 58.4        | 61.3        | 59.8        | 62.3        | 62.4        | 62.4        |
| OSCGA+TomBERT <sup>a</sup>    | 61.7        | 63.4        | 62.5        | 63.4        | 64.0        | 63.7        |
| UMT-collapsed <sup>a</sup>    | 60.4        | 61.6        | 61.0        | 60.0        | 61.7        | 60.8        |
| OSCGA-collapsed <sup>a</sup>  | 63.1        | 63.7        | 63.2        | 63.5        | 63.5        | 63.5        |
| RpBERT-collapsed <sup>a</sup> | 49.3        | 46.9        | 48.0        | 57.0        | 55.4        | 56.2        |
| JML <sup>a</sup>              | 65.0        | 63.2        | 64.1        | 66.5        | 65.5        | 66.0        |
| VLP-MABSA <sup>b</sup>        | 65.1        | 68.3        | <b>66.6</b> | 66.9        | 69.2        | 68.0        |
| CMMT <sup>c</sup>             | 64.6        | <b>68.7</b> | 66.5        | 67.6        | <b>69.4</b> | 68.5        |
| Our Model                     | <b>65.6</b> | 67.3        | 66.4        | <b>68.2</b> | 69.0        | <b>68.6</b> |

#### 4.5. Main results

In Table 2, the consistently superior performance of our underlying model, BART, in comparison to the other two methods, underscores its proficiency in tasks involving joint learning. For multimodal methods, previous pipeline approaches and collapsed tagging approaches perform much worse than recent joint learning approaches, probably because of the error propagation problem when these two subtasks are carried out separately. As the first joint learning method, JML performs better than previous studies since the joint framework improves the error propagation problem. Moreover, our model outperforms VLP-MABSA by 2.5% and 2.0%, with respect to the F1 score on TWITTER-15 and TWITTER-17, respectively. This is mainly benefits from its generative paradigm framework, which is superior in joint learning tasks. In conclusion, our proposed SAAF model distinctly attains the highest performance, as evaluated by the F1 score on the TWITTER-17 dataset. Furthermore, the F1 score of SAAF is only 0.2% lower on the TWITTER-15 dataset than VLP-MABSA which is highly pretrained. This demonstrates that SAAF is competitive among all the state of the art methods. These observations demonstrate the effectiveness of our SAAF model.

#### 4.6. Ablation study of adaptive visual-to-textual fusion layer

**Cross-modal interaction.** To verify the effect of cross-modal interaction, the unprocessed raw textual representation is directly input into both the selective fusion layer and the cross-modal mixup layer. The results are shown in Table 3.

**Table 3.** Ablation study of cross-modal interaction.

| Method                      | TWITTER-15  |             |             | TWITTER-17  |             |             |
|-----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                             | P           | R           | F1          | P           | R           | F1          |
| Our Model                   | <b>65.6</b> | <b>67.3</b> | <b>66.4</b> | <b>68.2</b> | <b>69.0</b> | <b>68.6</b> |
| w/o cross-modal interaction | 64.5        | 67.2        | 65.9        | 66.8        | 67.9        | 67.4        |

It can be seen that without cross-modal interaction, the F1 score of the TWITTER-15 and TWITTER-17 datasets drop by about 0.5% and 1.2%, respectively, compared to the full model. The above results further prove that extracting text closely related visual features can better achieve multi-modal fusion.

**Selective fusion.** Table 4 reports ablation study of the selective fusion layer. The unprocessed visual feature is directed into the cross-modal mixup layer instead of the fused representations. It can be seen that the performance drops sharply after the removal of selective fusion, illustrating the effectiveness of selective fusion layer, which aims to filter out unrelated region features for the text.

**Table 4.** Ablation study of selective fusion.

| Method               | TWITTER-15  |             |             | TWITTER-17  |             |             |
|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                      | P           | R           | F1          | P           | R           | F1          |
| Our Model            | 65.6        | <b>67.3</b> | <b>66.4</b> | <b>68.2</b> | <b>69.0</b> | <b>68.6</b> |
| w/o Selective Fusion | <b>66.5</b> | 63.5        | 65.0        | 66.0        | 67.6        | 66.8        |

**Cross-modal mixup.** The effectiveness of the cross-modal mixup layer is evaluated by omitting it from the adaptive visual-to-textual fusion layer. As can be seen in Table 5, the performance decreases

by 1.4% and 1.8% on the TWITTER-15 and TWITTER-17 datasets, respectively, after removing the cross-modal mixup layer, which illustrates the necessity of performing cross-modal mixup.

**Table 5.** Ablation study of cross-modal mixup.

| Method                | TWITTER-15  |             |             | TWITTER-17  |             |             |
|-----------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                       | P           | R           | F1          | P           | R           | F1          |
| Our Model             | 65.6        | 67.3        | <b>66.4</b> | <b>68.2</b> | <b>69.0</b> | <b>68.6</b> |
| w/o cross-modal mixup | <b>64.7</b> | <b>67.4</b> | 66.0        | 65.6        | 67.3        | 66.4        |

**Selective fusion & cross-modal mixup.** As can be seen in Table 6, w/o selective fusion & cross-modal mixup is the BART model only with our cross-modal interaction module. It performs worse after removing both the selective fusion layer and cross-modal mixup layer. It proves the effectiveness of the selective fusion layer and cross-modal mixup layer.

**Table 6.** Ablation study of selective fusion & cross-modal mixup.


| Method                                   | TWITTER-15  |             |             | TWITTER-17  |             |             |
|------------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                                          | P           | R           | F1          | P           | R           | F1          |
| Our Model                                | <b>65.6</b> | <b>67.3</b> | <b>66.4</b> | <b>68.2</b> | <b>69.0</b> | <b>68.6</b> |
| w/o Selective Fusion & Cross-Modal Mixup | 63.9        | 67.0        | 65.4        | 65.5        | 67.0        | 66.0        |

As indicated, the removal of either one or both modules (w/o selective fusion & cross-modal mixup) produce varying degrees of performance decline. This underscores the efficacy of the individual components, thereby augmenting the dependability and interpretability of our model.

4.7. Case study

To further demonstrate the effectiveness of our approach, we randomly select three samples from the TWITTER-17 dataset for a case study. Table 7 presents three test examples with predictions from two different baseline methods. The compared methods are Multimodal-BART (denoted by M-BART) and VLP. In the example (a), it is evident that both M-BART and VLP erroneously extracted the aspect term “Mott Basketball Camp.” In the example (b), M-BART failed to recognize the aspect term RutgersU, while VLP predicted the right aspect but wrongly predicted the sentiment toward the aspect term RutgersU as positive. Meanwhile, M-BART also failed to correctly predict the sentiment toward the aspect term Obama. For example (c), M-BART failed to extract the aspect term Pillers1957, while VLP extracted the wrong aspect term (i.e., KSC U10). However, among all cases, it is evident that our approach, SAAF, effectively extracts all aspect terms and accurately classifies sentiment by adaptively fusing visual and textual modalities for both subtasks within a generative framework.

**Table 7.** Predictions of M-BART, VLP and our model on three test samples. ✕ and ✓ denote incorrect and correct predictions.

|        |                                                                                   |                                                                                                           |                                                                                                                     |
|--------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Image  |  |                         |                                  |
| Text   | (a) Day 4 of Mott Basketball Camp.                                                | (b) Pres Obama takes the stage at @ RutgersU Commencement in school football stadium in Piscataway , NJ . | (c) Come support the KSC U10 Boys Soccer team BBQ 11 - 5 today . Thank You @ Pillers1957 for your generous donation |
| GT     | (Mott, POS)                                                                       | (Obama, POS)<br>(RutgersU, NEU)<br>(Piscataway, NEU)<br>(NJ, NEU)                                         | (KSC, POS)<br>(Pillers1957, NEU)                                                                                    |
| M-BART | (Mott Basketball Camp, NEU) ✕                                                     | (Obama, NEU) ✕<br>- ✕<br>(Piscataway, NEU) ✓<br>(NJ, NEU) ✓                                               | (KSC, POS) ✓<br>(11, POS) ✕                                                                                         |
| VLP    | (Mott Basketball Camp, POS) ✕                                                     | (Obama, POS) ✓<br>(RutgersU, POS) ✕<br>(Piscataway, NEU) ✓<br>(NJ, NEU) ✓                                 | (KSC U10, POS) ✕<br>(Pillers1957, NEU) ✓                                                                            |
| SAAF   | (Mott, POS) ✓                                                                     | (Obama, POS) ✓<br>(RutgersU, NEU) ✓<br>(Piscataway, NEU) ✓<br>(NJ, NEU) ✓                                 | (KSC, POS) ✓<br>(Pillers1957, NEU) ✓                                                                                |

## 5. Conclusions

In this paper, we propose self-adaptive cross-modal attention fusion architecture. This architecture leverages a selective fusion mechanism between image and text to bridge the semantic gap and enables the adaptive transfer of textual-based pre-training models to the multi-modal JMASA task. Experiment results show that our proposed approach generally outperforms many competitive unimodal and multimodal methods.

## Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Acknowledgments

This paper is supported by the Major Project of the Yunnan Provincial Science and Technology Department (202202AE090008-3), National Natural Science Foundation of China (62366025), Natural Science Foundation project of Yunnan Science and Technology Department (202301AT070444). Yunnan provincial major science and technology special plan projects (202103AA080015)

## Conflict of interest

The authors declare there is no conflict of interest.

## References

1. L. Zhu, M. Xu, Y. Bao, Y. Xu, X. Kong, Deep learning for aspect-based sentiment analysis: A review, *PeerJ Comput. Sci.*, **8** (2022), e1004. <https://doi.org/10.7717/peerj-cs.1044>
2. L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, *Inf. Fusion*, **95** (2023), 306–325. <https://doi.org/10.1016/j.inffus.2023.02.028>
3. J. Yu, J. Jiang, Adapting BERT for target-oriented multimodal sentiment classification, *IJCAI*, (2019), 5408–5414. <https://doi.org/10.24963/ijcai.2019/751>
4. Z. Khan, Y. Fu, Exploiting BERT for multimodal target sentiment classification through input space translation, in *Proceedings of the 29th ACM International Conference on Multimedia*, (2021), 3034–3042. <https://doi.org/10.1145/3474085.3475692>
5. J. Yu, K. Chen, R. Xia, Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis, *IEEE Trans. Affective Comput.*, **14** (2021), 1966–1978. <https://doi.org/10.1109/TAFFC.2022.3171091>
6. L. Yan, J. Yu, R. Xia, Vision-language pre-training for multimodal aspect-based sentiment analysis, preprint, arXiv: 2204.07955.
7. H. Xu, B. Liu, L. Shu, P. S. Yu, Double embeddings and CNN-based sequence labeling for aspect extraction, preprint, arXiv: 1805.04601.
8. P. Liu, S. Joty, H. Meng, Fine-grained opinion mining with recurrent neural networks and word embeddings, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (2015), 1433–1443. <https://doi.org/10.18653/v1/D15-1168>
9. D. Ma, S. Li, F. Wu, X. Xie, H. Wang, Exploring sequence-to-sequence learning in aspect term extraction, in *Proceedings of the 57th Annual Meeting of The Association for Computational Linguistics*, (2019), 3538–3547. <https://doi.org/10.18653/v1/P19-1344>
10. J. Yu, K. Chen, R. Xia, Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation, preprint, arXiv: 2004.14769.

11. C. Brun, D. N. Popa, C. Roux, XRCE: Hybrid classification for aspect-based sentiment analysis, in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, (2021), 838–842. <https://doi.org/10.3115/v1/S14-2149>
12. M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, et al., Semeval-2016 task 5: Aspect based sentiment analysis, in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, (2016), 19–30. <https://doi.org/10.18653/v1/S16-1002>
13. H. T. Nguyen, M. Le Nguyen, Effective attention networks for aspect-level sentiment classification, in *2018 10th International Conference on Knowledge and Systems Engineering (KSE), IEEE*, (2018), 25–30. <https://doi.org/10.1109/KSE.2018.8573324>
14. J. Cheng, S. Zhao, J. Zhang, I. King, X. Zhang, H. Wang, Aspect-level sentiment classification with heat (hierarchical attention) network, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, (2017), 97–106. <https://doi.org/10.1145/3132847.3133037>
15. J. Liu, Y. Zhang, Attention modeling for targeted sentiment, in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, **2** (2017), 572–577.
16. Y. Tay, L. A. Tuan, S. C. Hui, Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **32** (2018). <https://doi.org/10.1609/aaai.v32i1.12049>
17. Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (2016), 606–615.
18. X. Li, L. Bing, W. Lam, B. Shi, Transformation networks for target-oriented sentiment classification, preprint, arXiv: 1805.01086.
19. W. Xue, T. Li, Aspect based sentiment analysis with gated convolutional networks, preprint, arXiv: 1805.07043.
20. C. Chen, Z. Teng, Y. Zhang, Inducing target-specific latent structures for aspect sentiment classification, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2020), 5596–5607. <https://doi.org/10.18653/v1/2020.emnlp-main.451>
21. K. Wang, W. Shen, Y. Yang, X. Quan, R. Wang, Relational graph attention network for aspect-based sentiment analysis, preprint, arXiv: 2004.12362. <https://doi.org/10.48550/arXiv.2004.12362>
22. M. Zhang, T. Qian, Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2020), 3540–3549. <https://doi.org/10.18653/v1/2020.emnlp-main.286>
23. H. Tang, D. Ji, C. Li, Q. Zhou, Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification, in *Proceedings of the 58th Annual Meeting of The Association for Computational Linguistics*, (2020), 6578–6588. <https://doi.org/10.18653/v1/2020.acl-main.588>

24. B. Huang, K. M. Carley, Syntax-aware aspect level sentiment classification with graph attention networks, preprint, arXiv: 1909.02606. <https://doi.org/10.48550/arXiv.1909.02606>
25. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1** (2019), 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
26. R. Li, H. Chen, F. Feng, Z. Ma, X. Wang, E. Hovy, Dual graph convolutional networks for aspect-based sentiment analysis, in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, **1** (2021), 6319–6329. <https://doi.org/10.18653/v1/2021.acl-long.494>
27. M. Mitchell, J. Aguilar, T. Wilson, B. Van Durme, Open domain targeted sentiment, in *Proceedings of the 2013 conference on empirical methods in natural language processing*, (2013), 1643–1654.
28. M. Zhang, Y. Zhang, D. T. Vo, Neural networks for open domain targeted sentiment, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (2015), 612–621. <https://doi.org/10.18653/v1/D15-1073>
29. M. Hu, Y. Peng, Z. Huang, D. Li, Y. Lv, Open-domain targeted sentiment analysis via span-based extraction and classification, preprint, arXiv: 1906.03820.
30. H. Yan, J. Dai, X. Qiu, Z. Zhang, A unified generative framework for aspect-based sentiment analysis, preprint, arXiv: 2106.04300.
31. N. Xu, W. Mao, G. Cheng, Multi-interactive memory network for aspect based multimodal sentiment analysis, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 371–378. <https://doi.org/10.1609/aaai.v33i01.3301371>
32. D. Gu, J. Wang, S. Cai, C. Yang, Z. Song, H. Zhao, et al., Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network, *IEEE Access*, **9** (2021), 157329–157336. <https://doi.org/10.1109/ACCESS.2021.3126782>
33. J. Yu, J. Jiang, R. Xia, Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification, *IEEE/ACM Trans. Audio Speech Language Process.*, **28** (2019), 429–439. <https://doi.org/10.1109/TASLP.2019.2957872>
34. Z. Zhang, Z. Wang, X. Li, N. Liu, B. Guo, Z. Yu, ModalNet: An aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network, *World Wide Web*, **24** (2021), 1957–1974. <https://doi.org/10.1007/s11280-021-00955-7>
35. X. Ju, D. Zhang, R. Xiao, J. Li, S. Li, M. Zhang, et al., Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (2021), 4395–4405. <https://doi.org/10.18653/v1/2021.emnlp-main.360>
36. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, preprint, arXiv: 1910.13461.

37. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, et al., Bottom-up and top-down attention for image captioning and visual question answering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 6077–6086.
38. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.*, **30** (2017).
39. J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, preprint, arXiv: 1607.06450.
40. Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, R. Salakhutdinov, Attention is all you need, in *Proceedings of the conference Association for Computational Linguistics. Meeting*, **2019** (2019), 6558–6569. <https://doi.org/10.18653/v1/p19-1656>
41. J. Chen, Z. Yang, D. Yang, Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification, preprint, arXiv: 2004.12239.
42. Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, R. Salakhutdinov, Km-bart: Knowledge enhanced multimodal bart for visual commonsense generation, preprint, arXiv: 2101.00419.
43. G. Chen, Y. Tian, Y. Song, Joint aspect extraction and sentiment analysis with directional graph convolutional networks, in *Proceedings of the 28th International Conference on Computational Linguistics*, (2020), 272–279. <https://doi.org/10.18653/v1/2020.coling-main.24>
44. J. Yu, J. Jiang, L. Yang, R. Xia, Improving multimodal named entity recognition via entity span detection with unified multimodal transformer, *Assoc. Comput. Linguist.*, (2020), 272–279.
45. H. Wu, S. Cheng, J. Wang, S. Li, L. Chi, Association for Computational Linguistics, in *Natural Language Processing and Chinese Computing: 9th CCF International Conference*, (2020), 145–156.
46. L. Sun, J. Wang, K. Zhang, Y. Su, F. Weng, RpBERT: A text-image relation propagation-based BERT model for multimodal NER, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 13860–13868. <https://doi.org/10.1609/aaai.v35i15.17633>
47. L. Yang, J. C. Na, J. Yu, Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis, *Inf. Process. Manage.*, **59** (2022), 103038. <https://doi.org/10.1016/j.ipm.2022.103038>



AIMS Press

©2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)