Knowledge-Based Systems 279 (2023) 110951

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Collaborative fine-grained interaction learning for image–text sentiment analysis

Xingwang Xiao [a], Yuanyuan Pu [a,b,*], Dongming Zhou [a], Jinde Cao [c], Jinjing Gu [a], Zhengpeng Zhao [a], Dan Xu [a]

[a] College of Information Science and Engineering, Yunnan University, Kunming 650500, China
[b] University Key Laboratory of Internet of Things Technology and Application, Yunnan Province, Kunming 650500, China
[c] College of Automation, Southeast University, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 10 June 2023
Received in revised form 1 August 2023
Accepted 30 August 2023
Available online 4 September 2023

Keywords:
Image–text sentiment analysis
Fine-grained interaction
Image–text dataset
Memory transformer

ABSTRACT

Investigating interactions between image and text can effectively improve image–text sentiment analysis, but most existing methods do not explore image–text interaction at fine-grained level. In this paper, we propose a Memory-enhanced Collaborative Fine-grained Interaction Transformer (MCFIT) to learn collaborative fine-grained interaction between image and text. Specifically, a multi-branch encoder is designed to learn both fine-grained region-word and patch-word interactions. Meanwhile, Memory-enhanced Cross-Attention (MECA) is proposed to utilize patch and region information to improve region-word interaction and patch-word interaction learning, respectively. Therefore, collaborative fine-grained interaction can yield more accurate image–text interaction. Finally, to analyze the sentiments embedded in real-life Chinese image–text pairs, we build a large-scale Chinese image–text sentiment dataset (CISD) containing 54,931 image–text pairs. Extensive experiments conducted on four real-life datasets prove the effectiveness of collaborative fine-grained interaction and demonstrate that MCFIT outperforms the state-of-the-art baselines.

1. Introduction

As the use of mobile terminal devices enters a period of explosive growth, more and more users trend toward posting single-modal and multimodal data (e.g., image, text, and video) to convey their emotions and opinions on different social media platforms, such as Twitter, Weibo, and Tumblr. Research with the purpose of analyzing the sentiments embedded in single-modal and multimodal data has received increasing attention in recent years. Besides, compared to single-modal data, multi-modal data possesses more information and makes the user's expression more vivid. Therefore, exploring multimodal sentiments has many applications, including multimodal sentiment analysis [1–3], decision making [4], cross-modal information retrieval [5, 6] and so on.

For multimodal sentiment analysis, we focus on image–text sentiment analysis on real-life image–text pairs. With the development of deep learning, image–text sentiment analysis based on deep learning has made exciting progress in recent years. Most previous studies simply concatenate features extracted from different modalities [7–10], which could not sufficiently explore the relationship between image and text. Meanwhile, some works focus on modeling the relationship and interaction of image–text pair [11–14]. However, they are not fine-grained interaction methods.

Typically, salient affective regions of an image are able to evoke the sentiments of human [15–18] and some words in the comment for an image are usually related to the affective regions [19]. Therefore, as shown in A.1 of Fig. 1, some works [19–21] have studied the interaction between regions and words and have proved that fine-grained region-word interaction can improve the performance of sentiment analysis. However, the affective regions (boxed in red) are detected by pre-trained Faster R-CNN [22] and the corresponding non-affective regions (filled in green) are discarded, which will result in the loss of the necessary information embedded in the non-affective regions. Specifically, as shown in B.1, the affective region will not be able to get correct interaction with the words in the text because the affective region (**smiling man**) is positive and the words (**He is not doing well at all**) are negative. Obviously, in similar cases, the affective regions would not represent the complete sentiment information of the images.

* Corresponding author.
E-mail addresses: xiaoxingwang_0612@163.com (X. Xiao), yuanyuanpu@ynu.edu.cn (Y. Pu), zhoudm@ynu.edu.cn (D. Zhou), jdcao@seu.edu.cn (J. Cao), jinjinggu@ynu.edu.cn (J. Gu), zhpzhao@ynu.edu.cn (Z. Zhao), danxu@ynu.edu.cn (D. Xu).

https://doi.org/10.1016/j.knosys.2023.110951
0950-7051/

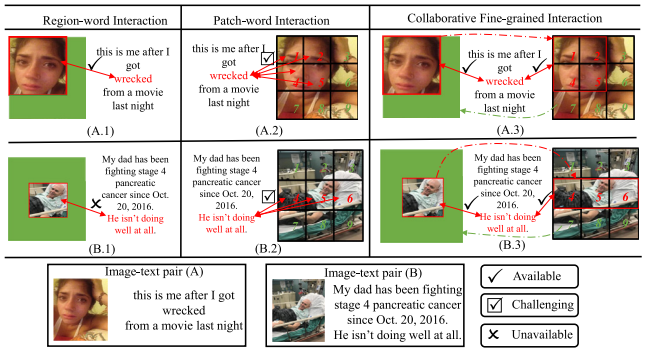

Fig. 1. For region-word interaction, non-affective regions are discarded, which loses the auxiliary information embedded in them. Besides, there are more affective patches than regions, which makes the patch-word interaction more challenging. In collaborative fine-grained interaction, region with bounding box information is employed to help patch-word interaction locate affective patches and the information embedded in the non-affective patches is utilized to improve region-word interaction learning.

As Vision Transformer (ViT) [23] states, an image is worth 16 × 16 words (patches). A fusion method based on Cross-Attention [24] is proposed by [25] to model the interaction between patch-level image features and word-level text features. This fine-grained interaction between patches and words (patch-word interaction) also yields promising performance for sentiment analysis. As illustrated in Fig. 1 (A.2 and B.2), since both affective patches (marked by red numbers) and non-affective patches (marked by green numbers) are preserved, the information embedded in the non-affective patches will be utilized. In other words, the patches can represent the complete sentiment information of the images. Accordingly, the affective patches can interact well with the corresponding words. For example, compared to B.1, B.2 is able to obtain the correct interaction. However, since the affective patches are more than affective regions, patch-word interaction is more challenging than region-word interaction. For instance, as shown in A.1 and A.2, four patches need to interact with one word, which would be more difficult than one region interacting with one word.

Overall, if we can utilize the advantages of both region-word and patch-word interactions, we will obtain a better sentiment analysis performance. More specifically, as demonstrated in A.3 and B.3, we employ region with bounding box information to help patch-word interaction locate affective patches and utilize the information embedded in the non-affective patches to improve region-word interaction learning.

To achieve the above goals, in this paper, we propose a Memory-enhanced Collaborative Fine-grained Interaction Transformer (MCFIT). In MCFIT, we design a multi-branch encoder to learn both fine-grained region-word and patch-word interactions. Besides, Memory-enhanced Cross-Attention (MECA), which can improve region-word interaction learning with patch information and facilitate patch-word interaction learning with region information, is introduced. Finally, we observe that the publicly available real-life image–text datasets employed for image–text sentiment analysis are all in English (e.g., MVSA [26] and TumEmo [12]). In other words, publicly available real-life Chinese image–text dataset is anxious to be reported. Thus, to analyze the sentiments embedded in Chinese image–text pairs, we build a large-scale Chinese image–text sentiment dataset (CISD) including 54,931 Chinese image–text pairs.

Our main contributions can be summarized as follows:

- A large-scale Chinese image–text dataset including 54,931 real-life Chinese image–text pairs is reported, which allows us to analyze the sentiments embedded in Chinese image–text pairs via the proposed method.
- We propose a novel MCFIT to achieve image–text sentiment analysis by learning collaborative fine-grained interaction between image and text. In MCFIT, both fine-grained region-word and patch-word interactions are employed to improve the performance of sentiment analysis. Meanwhile, a MECA module is introduced, which leverages region information and patch information to help patch-word interaction and region-word interaction learning, respectively.
- Extensive experiments are conducted on four real-life image–text datasets, and the results demonstrate that the proposed model outperforms state-of-the-art methods.

2. Related work

2.1. Image–text sentiment analysis

In the early years, traditional machine learning methods were employed for image–text sentiment analysis [27,28]. With the development of deep learning, methods based on deep learning have achieved surprising performance in recent years.

Xu et al. developed three methods including HSAN [7], Co-MN [11], and MultiSenti [29] to achieve image–text sentiment analysis. To explore the structure of emotions from image–text pairs, Hu et al. [8] concatenated text and image features. For MeToo image–text data, Basu et al. [9] constructed a network to concatenate image and text features for analyzing the sentiments. Yang et al. proposed two methods including MVAN [12] and MGNNS [13] for image–text sentiment analysis. Among them, a image–text emotion dataset was reported in [12]. Following MGNNS [13], graph neural network was applied by [30] to model the interaction between image and text. Zhang et al. [31] proposed a disentangled sentiment representation adversarial network to reduce the domain shift of expressive styles for multimodal cross-domain sentiment analysis. Yu et al. [32] developed a hierarchical interactive multimodal transformer to achieve aspect-based image–text sentiment analysis. Li et al. [33] constructed a novel contrastive learning and multi-layer fusion method for multimodal sentiment detection. Hu et al. [34] presented a two-stage attention-based neural network to analyze textual-visual content for sentiment classification. Self-supervised learning was employed by Zhu et al. [35] to achieve image–text sentiment analysis via image–text matching. Liu et al. [36] proposed a model via scanning, attention, and reasoning image–text for sentiment analysis. Similarly, Pandey et al. [21] employed spatio-depth visual attention for visual-caption sentiment recognition.

Since salient affective regions of an image are able to evoke the sentiments of human, Zhu et al. [19] constructed a novel image–text interaction network to investigate the relationship between affective image regions and text words for image–text sentiment analysis. Similarly, Du et al. [37] proposed a gated attention fusion network to investigate the relationship between affective image regions and text words. Furthermore, Liang et al. [38] introduced a graph-based fusion method to model the relationship between regions and words.

Although current image–text sentiment analysis methods have achieved promising results, most of the approaches perform simple fusion of image and text features or model their interaction at a shallow level. In this paper, a collaborative fine-grained interaction between image and text is proposed, which can model fine-grained region-word and patch-word interactions.

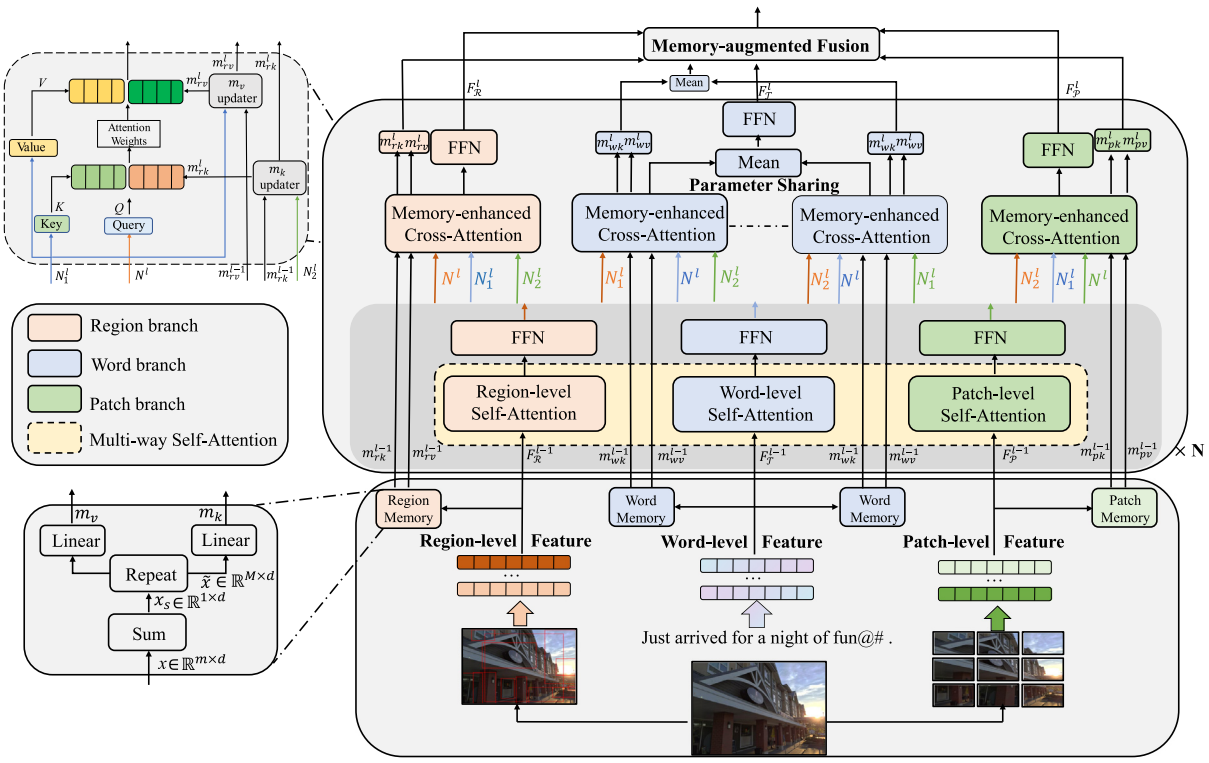

Fig. 2. The framework of Memory-enhanced Collaborative Fine-grained Interaction Transformer. For an image–text, text is encoded to word-level feature, image is encoded to region-level and patch-level features. Then, we design a multi-branch encoder to achieve collaborative fine-grained interaction shown in Fig. 1 (A.3 and B.3). In this model, patch information and region information are utilized to improve region-word interaction and patch-word interaction learning via Memory-enhanced Cross-Attention demonstrated in Fig. 3, respectively. Finally, as shown in Fig. 4, we design a Memory-augmented Fusion Network to exploit the information embedded in the memory vectors.

2.2. Transformer with memory

Memory has a fundamental role in human cognition. Humans neurocode perceptual information and store it in memory, and then our brains can retrieve this information to help us solve relevant tasks efficiently [39]. Many studies about modeling the behavior of human memory have been applied to artificial neural networks. In early years, some works including RNN [40], LSTM [41] and GRU [42] modeled temporal sequences by considering their internal state vector as single memory slot. However, limited by memory bottleneck, they are unable to store long-term information.

More recently, Transformer [24] and its followers have achieved great success in NLP by using multi-head self-attention to model sequence dependencies rather than the use of memory. However, when the input sequence is long, self-attention consumes a large amount of computation. Transformer-XL [43] re-employed the concept of memory for Transformer to model the long-term dependencies in long sequences. Since then, Transformer with memory has achieved significant success. Adel et al. [44] employed memory Transformer and hierarchical attention to process long documents. Lei et al. [45] introduced a memory-augment Transformer for video paragraph captioning. Similarly, an enhanced-memory Transformer was proposed by [46] to achieve video paragraph captioning. Cornia et al. [47] constructed a meshed-memory Transformer for image captioning.

We observe that it is a fact that memory enhances the performance of Transformer, and there is no memory-related Transformer study for image–text sentiment analysis. Motivated by this, we propose a memory Transformer with Memory-enhanced Cross-Attention to explore collaborative fine-grained interaction between image and text.

3. Methodology

In this section, we elaborate on the details of the proposed Memory-enhanced Collaborative Fine-grained Interaction Transformer for image–text sentiment analysis. The architecture of MCFIT is shown in Fig. 2.

3.1. Feature encoding

Region-level Feature. To represent an image $\boldsymbol{I}$ with regions, following [48], we detect image regions and their associated representations utilizing Faster R-CNN [22], which is pre-trained on Visual Genomes dataset [49] using ResNet-101 [50] as backbone. In our work, top-m regions are selected and each region becomes a 2048-dimensional feature vector after average-pooling operation, which is defined as $f_i (i = 1, 2, \ldots, m)$. We then project each vector to a d-dimensional region feature r_i. Absolute positional encoding (APE) tells the model where the feature is, which is important information. Suppose there are two objects with identical region features: one locates in the corner and the other locates in the center. In this case, APE facilitates the model to distinguish them accurately. Thus, we add APE like [24] to all regions:

$$r_i = \mathbf{PE}_i + (W_r f_i + b_r), i \in [1, m], \tag{1}$$

where W_r and b_r are learnable parameters, and $\mathbf{PE}_i$ is the position encoding of ith region feature.

Patch-level Feature. As stated in ViT, an image is worth 16×16 words (patches). Thus, to represent image patch-level features, we apply the ViT stream of CLIP (ViT-B/16) [51] that are pre-trained on 400 million image–text pairs and remove the last

[**CLS**] features choosing operation to embed each image patch to 512-dimensional patch-level feature $p_i (i = 1, 2, \ldots, p)$.

Word-level Feature. To represent a sentence $\boldsymbol{T}$ with k words, we employ the Transformer stream of CLIP (ViT-B/16) [51] and remove the last [**EOT**] features choosing operation. Finally, each word of the sentence is embedded to 512-dimensional word-level feature $w_i (i = 1, 2, \ldots, k)$.

In general, given an image–text pair, it is encoded to a region-level feature set $\mathcal{R} = \{r_1, r_2, \ldots, r_m\}$, a patch-level feature set $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, and a word-level feature set $\mathcal{T} = \{w_1, w_2, \ldots, w_k\}$. Meanwhile, as shown in Fig. 2, three feature sets are sent to memory initializers to get initialized memory vectors.

$$\begin{aligned} x_s &= \text{Sum}\,(x)\,, (x_s \in \mathbb{R}^{1\times d}), \\ \tilde{x_s} &= \text{Repeat}\,(x_s)\,, (\tilde{x_s} \in \mathbb{R}^{M\times d}), \\ m_k &= W_k \tilde{x_s} + b_k, (m_k \in \mathbb{R}^{M\times d}), \\ m_v &= W_v \tilde{x_s} + b_v, (m_v \in \mathbb{R}^{M\times d}), \end{aligned} \tag{2}$$

where, W_k, W_v, b_v, and b_k are learnable weights. M is the length of memory vectors, which is analyzed in Section 4.8.

3.2. Multi-branch Memory-enhanced encoder

We elaborately construct a stackable multi-branch encoder to model fine-grained region-word and patch-word interactions between image and text. The encoder consists of two sub-modules: Multi-Way Self-Attention (MWSA) and Memory-enhance Cross-Attention (MECA). Among them, MWSA is to model the intra-level relationship and MECA is to realize two fine-grained interactions. In MECA, the patch information and region information are utilized to enhance the region-word interaction and patch-word interaction learning, respectively.

Multi-Way Self-Attention. In order to model intra-level relationships of three kinds of features to improve cross-modal interaction learning, we devise a Multi-Way Self-Attention (MWSA) which consists of three independent self-attention modules.

Word-level Self-Attention. We employ Multi-Head Self-Attention (MHSA) [24] to further model the information between words. In lth layer, MHSA is defined as follows:

$$\begin{aligned} \text{SA} &= \text{Attention}\,(Q, K, V) \\ &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V, \\ \text{MHSA} &= \text{Concat}(\text{head}_1, \ldots, \text{head}_n) W_{\mathcal{T}}, \\ \text{where, head}_i &= \text{SA}\left(W_q^i F_{\mathcal{R}}^l, W_k^i F_{\mathcal{R}}^l, W_v^i F_{\mathcal{R}}^l\right), \end{aligned} \tag{3}$$

where W_q^i, W_k^i, W_v^i, $W_{\mathcal{T}}$ are matrices of learnable weights.

Patch-level Self-Attention. To further model intra-level relationships of patch-level features, following [52], image Relative Position Encoding (iRPE) for ViT is employed to further learn the contextual information between each p_i of the input patch-level feature set $\mathcal{P}$. In lth layer, Multi-Head Patch-level Self-Attention (MHPSA) is described as:

$$\begin{aligned} \text{PSA} &= \text{Attention}(Q, K, V, \Omega) = \\ &\text{softmax}\left(\frac{QK^T + \text{iRPE}}{\sqrt{d}}\right) V, \\ &\text{where, iRPE} = K\Omega^T, \\ \text{MHPSA} &= \text{Concat}\,(\text{head}_1, \ldots, \text{head}_n)\, W_{\mathcal{C}}, \\ \text{where, head}_i &= \text{PSA}\left(W_q^i F_{\mathcal{P}}^l, W_k^i F_{\mathcal{P}}^l, W_v^i F_{\mathcal{P}}^l, \Omega_i\right), \end{aligned} \tag{4}$$

where W_q^i, W_k^i, W_v^i, W_g, Ω_i are matrices of learnable weights of the ith head.

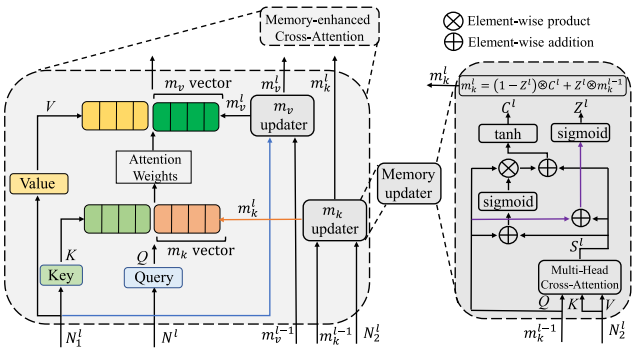

Fig. 3. The framework of Memory-enhanced Cross-Attention. m_x^l $(x \in [v, k])$ that contains the information of m_x^{l-1} and N_y^l $(y \in [1, 2])$ will contribute to obtaining accurate interactions in lth layer. Thus, patch and region information can be employed to improve region-word interaction and patch-word interaction learning via MECA, respectively.

Region-level Self-Attention. Generally, region-level features are extracted by locally-connected convolutions, which make regions relation-weak and even isolated. To solve this issue, we also apply self-attention to model and enhance the relationships between the regions. Besides, the bounding box of a region can be represented as (x, y, w, h) where x, y, w, and h denote the box's center coordinates and its width and height. For the geometry features of box_n and box_m, their geometry relation representation can be denoted as a 4-d vector. Inspired by [53], to integrate relative location information of region features, the geometric Relative Position Encoding (gRPE) is adopted. Similarly, iRPE is also applied and Multi-Head Region-level Self-Attention (MHRSA) is defined as:

$$\begin{aligned} \text{RSA} &= \text{Attention}\,(Q, K, V, \Omega, \varepsilon_{\mathcal{R}}) \\ &= \text{softmax}\left(\frac{QK^T + \text{iRPE}}{\sqrt{d}} + \text{gRPE}\right), \\ \text{where, iRPE} &= K\Omega^T, \text{gRPE} = \log\left(\varepsilon_{\mathcal{R}}\right), \\ \text{MHRSA} &= \text{Concat}\,(\text{head}_1, \ldots, \text{head}_n)\, W_{\mathcal{R}}, \\ \text{where, head}_i &= \text{RSA}\left(W_q^i F_{\mathcal{R}}^l, W_k^i F_{\mathcal{R}}^l, W_v^i F_{\mathcal{R}}^l, \Omega_i, \varepsilon_{\mathcal{R}}\right), \end{aligned} \tag{5}$$

where W_q^i, W_k^i, W_v^i, $W_{\mathcal{R}}$, Ω_i are matrices of learnable weights. The $\varepsilon_{\mathcal{R}}$ is defined as: $\varepsilon_{\mathcal{R}} =$

$$\left(\log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{|y_m - y_n|}{h_m}\right), \log\left(\frac{w_n}{w_m}\right), \log\left(\frac{h_n}{h_m}\right)\right)^T \tag{6}$$

Then we adopt three independent position-wise feedforward networks FFN for each branch features (f_s) encoded by Self-Attention modules.

$$\tilde{f_s} = W_{s2}\,(f_s W_{s1} + b_{s1}) + b_{s2}, \tag{7}$$

where $s \in [\mathcal{R}, \mathcal{T}, \mathcal{P}]$ and W_{s1}, W_{s2} are learnable weights, b_{s1} and b_{s2} are learnable biases. After that, the relation-aware representations are fed into next module.

Memory-enhanced Cross-Attention. Generally, Cross-Attention is utilized to fuse different modalities or get interaction between two modalities. Therefore, Cross-Attention can be employed to achieve fine-grained path-word and region-word interactions as shown in Fig. 1 (A.1 and A.2). However, as illustrated in Fig. 1 (A.3 and B.3), since we need to leverage the information of region and patch to enhance the region-word interaction and patch-word interaction learning respectively, vanilla Cross-Attention is inapplicable.

To address this challenge, following [47] that employs memory vectors to store additional information, we propose a novel Memory-enhanced Cross-Attention, which is visualized in Fig. 3. In MECA, the sets of keys and values used for Cross-Attention are extended with two additional memory vectors (m_k and $m_v \in \mathbb{R}^{M\times d}$) that can store the information from regions or patches. In this way, patch and region information can be utilized to improve region-word interaction and patch-word interaction learning, respectively. Formally, Multi-Head Memory-enhanced Cross-Attention (MHMECA) is defined as:

$$\begin{aligned} \text{MECA} &= \text{Attention}\left(Q, K_m, V_m\right), \\ \text{where, } K_m &= [K, m_k] \text{ and } V_m = [V, m_v], \\ \text{MHMECA} &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_n\right) W_m, \\ \text{where, head}_i &= \text{MECA}\left(W_q^i N, [W_k^i N_1, m_k], [W_v^i N_1, m_v]\right), \end{aligned} \tag{8}$$

where W_q^i, W_k^i, W_v^i and W_m are matrices of learnable weights and [,] indicates concatenation. Note that, m_k and m_v will be updated by their own memory updater before concatenating with K and V. Specifically, assume that m_k^l and m_v^l are the memories of lth layer in our stacked encoder, m_k^l is updated by m_k^{l-1} and N_2^l, m_v^l is updated by m_v^{l-1} and N_1^l. We summarize the update procedure as below:

$$\begin{aligned} &\text{Cross-Attention} = \text{CA} = \text{Attention}\left(Q, K, V\right), \\ &S^l = \text{MHCA} = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_n\right) W_U, \\ &\text{where, head}_i = \text{CA}\left(W_q^i m_x^{l-1}, W_k^i N_y^l, W_v^i N_y^l\right), \\ &R^l = \text{sigmoid}\left(W_{mr}^l m_x^{l-1} + W_{sr}^l S^l + b_r^l\right), \\ &Z^l = \text{sigmoid}\left(W_{mz}^l m_x^{l-1} + W_{sz}^l S^l + b_z^l\right), \\ &C^l = \tanh\left(R^l \otimes W_{mc}^l m_x^{l-1} + W_{sc}^l S^l + b_c^l\right), \\ &m_x^l = \left(1 - Z^l\right) \otimes C^l + Z^l \otimes m_x^{l-1}, \end{aligned} \tag{9}$$

where $x \in [v, k]$ and $y \in [1, 2]$ are paired in order, $\otimes$ denotes element-wise product, and MHCA is Multi-Head Cross-Attention [24]. W_q^i, W_k^i, W_v^i, W_U, W_{mr}^l, W_{sr}^l, W_{mz}^l, W_{sz}^l, W_{mc}^l and W_{sc}^l are learnable weights, b_r^l, b_z^l and b_c^l are learnable biases. C^l is the internal cell state, R^l is the reset gate that controls how much information from the previous memory state m_x^{l-1} needs to be forgotten. Z^l is the update gate that determines how much information from the previous memory state m_x^{l-1} as well as the current state N_y^l needs to be preserved. Clearly, m_x^l that contains the information from m_x^{l-1} and N_y^l will contribute to obtaining accurate region-word and patch-word interactions in lth layer. Note that this update strategy (memory updater) is conceptually similar to GRU [42].

As illustrated in Fig. 2, for the region branch, N^l, N_1^l and N_2^l represent region-level features $\mathcal{R}$, word-level features $\mathcal{T}$ and patch-level features $\mathcal{P}$ respectively. Since m_k and m_v of region branch are obtained by compressing the region-level features following Eq. (2), they contain the information of regions. And meanwhile, before concatenating with K and V, they are updated by N_2 and N_1, respectively. As shown in Fig. 1 A.3, K containing information from non-affective patches will enhance the region-word interaction learning. Besides, V including information from words will provide more comprehensive word information, which also further improves the region-word interaction learning. Similarly, for the patch branch, m_k updated by region features that contain bounding box position information will help to locate relevant affective patches during modeling patch-word interaction, which effectively reduces the challenge of patch-word interaction learning. And V including information from words can further improves the accuracy of patch-word interaction learning. For the word branch, text interacts with dual-level features of image in two parameter sharing MECAs. Finally, in order to stack the encoder to explore deep sentiment interactions, we take averaging operation to keep features dimension invariable. Then we also adopt three FFNs for each branch features (f_m) encoded by MECA.

$$F_m^l = W_{m2}\left(f_m W_{m1} + b_{m1}\right) + b_{m2}, \tag{10}$$

where $m \in [\mathcal{R}, \mathcal{T}, \mathcal{P}]$ and W_{m1}, W_{m2} are learnable weights, b_{m1} and b_{m2} are learnable biases.

Overall, this encoder enables fine-grained region-word and patch-word interaction learning. As shown in Fig A.3 and B.3, with this encoder, the patch information and region information are utilized to enhance the region-word interaction and patch-word interaction learning, respectively. Therefore, our collaborative fine-grained interaction can yield accurate image–text interaction, which will improve the performance of image–text sentiment analysis.

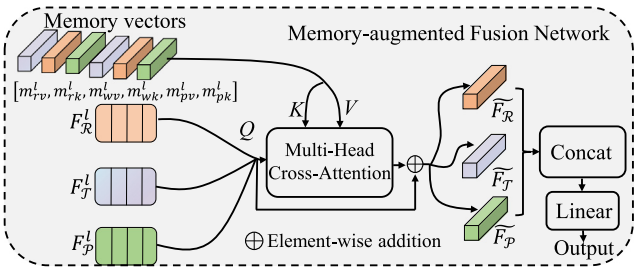

Fig. 4. The architecture of Memory-augmented Fusion Network. In this network, the information embedded in memory vectors is further utilized to improve the performance of sentiment analysis.

3.3. Memory-augmented Fusion Network

Since memory vectors are dynamically updated layer by layer in stacked encoder, necessary sentiment interaction information will be embedded in the output memory vectors. To fully leverage the information gathered in the memory vectors, we considerately design a Memory-augmented Fusion Network.

As demonstrated in Fig. 4, the sentiment-aware representations are used sequentially as query to obtain the sentimental information in concatenated memory vectors ($m_{all} \in \mathbb{R}^{6M\times d}$), and then respectively add to the corresponding sentiment-aware representation. Finally, the memory-augmented sentiment-aware representations are concatenated and fed into a single-linear layer. Note that, for the finally memory vectors from two parameter sharing MECAs, averaging operation is utilized. We summarize this procedure as:

$$\begin{aligned} m_{\text{all}} &= \left[m_{rv}^l, m_{rk}^l, m_{wv}^l, m_{wk}^l, m_{pv}^l, m_{pk}^l\right], \\ \tilde{F}_{\mathcal{R}} &= F_{\mathcal{R}}^l + \text{MHCA}\left(F_{\mathcal{R}}^l, m_{\text{all}}, m_{\text{all}}\right), \\ \tilde{F}_{\mathcal{T}} &= F_{\mathcal{T}}^l + \text{MHCA}\left(F_{\mathcal{T}}^l, m_{\text{all}}, m_{\text{all}}\right), \\ \tilde{F}_{\mathcal{P}} &= F_{\mathcal{P}}^l + \text{MHCA}\left(F_{\mathcal{P}}^l, m_{\text{all}}, m_{\text{all}}\right), \\ \tilde{F}_o &= W_o\left(\left[\tilde{F}_{\mathcal{R}}, \tilde{F}_{\mathcal{T}}, \tilde{F}_{\mathcal{P}}\right]\right) + b_o, \end{aligned} \tag{11}$$

where [,] means concatenation, W_o and b_o are learnable weight and bias. F_x and $\tilde{F}_x (x \in [\mathcal{R}, \mathcal{T}, \mathcal{P}])$ represent sentiment-aware and memory-augmented sentiment-aware representation. $m_z^l (z \in [rv, rk, wv, wk, pv, pk])$ is the final layer memory vectors of three branches.

3.4. Image–text sentiment classification

The objective of our model is to identify which sentiment is conveyed by an image–text input. Thus, we feed the fused

memory-augmented sentiment-aware representation $\tilde{F}_o$ into softmax function for image–text sentiment classification. Besides, our model is optimized by minimizing the cross-entropy loss.

$$\hat{y} = \text{softmax}\left(\tilde{F}_o\right),$$
$$\mathcal{L} = -\sum_i y_i \log \hat{y}_i, \tag{12}$$

where y_i denotes the ground truth sentiment label, and $\hat{y}_i$ is the output of the softmax.

4. Experiments

4.1. Datasets

We observe that although there are publicly available real-life image–text datasets including MVSA-Single (MVSA-S), MVSA-Multiple (MVSA-M) [26], and TumEmo [12], they are all in English. In other words, no real-life Chinese image–text datasets has been reported.

In this paper, to analyze the sentiments embedded in Chinese image–text pairs, we collect a large number of real-life Chinese image–text pairs from Wukong dataset [54] and Weibo. Specifically, 50,000 and 10,000 image–text pairs are collected from Wukong[1] and Weibo,[2] respectively. On the other hand, since fully manual annotation is very time-consuming and laborious, following [12], distant supervision is employed. Firstly, we randomly select and manually label 600 positive image–text pairs, 600 negative image–text pairs, and 600 neutral image–text pairs from the collected 60,000 image–text pairs. Secondly, the feature mean (FM) and maximum Euclidean distance (MED) are calculated from the positive, negative, and neutral image–text features, respectively. For instance, positive FM represents the mean of the positive image–text features, positive MED represent the maximum Euclidean distance between the positive image–text features and FM. Finally, we employ FM as supervisor to classify the unlabeled image–text pairs. Specifically, when the Euclidean distance between an unlabeled image–text pair and the positive FM is less than positive MED, this image–text pair will be labeled as positive. Note that, if the Euclidean distance is larger than MED, this image–text pair will be removed. Eventually, a distant-supervised Chinese image–text sentiment dataset (CISD[3]) including 54,931 Chinese image–text pairs is built, and some samples are reported in Fig. 5.

The proposed model is evaluated on four real-life datasets including MVSA-S, MVSA-M [26], TumEmo [12], and CISD. MVSA-S and MVSA-M are two different scale image–text sentiment datasets collected from Twitter. TumEmo is an image–text weak-supervision emotion dataset that contains large-scale image–text pairs collected from Tumblr. The statistics of four datasets used for our experiments are listed in Table 1.

4.2. Implementation details

In our experiments, four datasets are divided into training set, validation set and test set at a ratio of 8:1:1. We utilize Adam to optimize our model that is implemented by PyTorch. Top-m regions are 36 on all datasets. Batch sizes of MVSA-S, MVSA-M, TumEmo, and CISD are set to 16, 32, 32, and 32 respectively. The initial learning rate is 1e−4. To better train our model, the learning rate scheduler StepLR is employed with different settings on different dataset, where step size and gama are 1 and 0.8 on MVSA-S and MVSA-M, step size and gama are 2 and 0.8 on TumEmo and CISD. In our model, d-dimension and head number are set to 512 and 8.

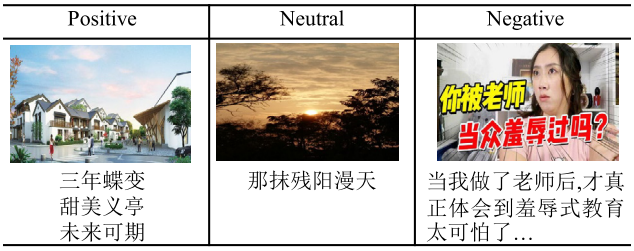

Fig. 5. Some Chinese image–text pairs of CISD.

Table 1
The statistics of four datasets.

	Positive	Neutral	Negative	All
MVSA-S	2,683	470	1,358	4,511
MVSA-M	11,318	4,408	1,298	17,024
CISD	13,598	24,869	16,464	54,931
		TumEmo		
Angry			14,554	
Bored			32,283	
Calm			18,109	
Fear			20,264	
Happy			50,267	
Cove			34,511	
Sad			25,277	
All			195,265	

Evaluation metrics. Evaluation metrics including accuracy **(Acc)** and F1-score **(F1)** are employed to evaluate all models.

4.3. Image–text sentiment analysis baselines

We compare our model with the following image–text sentiment analysis baselines.

MMBT [55] is a multimodal bidirectional transformer reported by Facebook AI Research for image–text classification.

MMBT-F [9] applies MMBT to achieve sentiment analysis of imbalanced image–text datasets.

D-Senti [8] concatenates visual and textual features to predict sentiment label.

DMAF [14] is a deep attentive fusion method for image–text sentiment analysis.

Co-MN [11] is a co-memory network for iteratively modeling the relations between image and text.

MVAN [12] is a multi-view attentional network that utilizes a memory network for image–text emotion analysis.

MGNNS [13] introduces a multi-channel graph neural networks to learn image–text representations based on the global characteristics of the dataset.

TSAFNet [34] presents a two-stage attention-based neural network to analyze textual-visual content for sentiment classification.

CLMLF [33] employs two contrastive learning tasks (label based contrastive learning and data based contrastive learning tasks) to help the model learn common features related to sentiment in image–text data.

ITIN [19] proposes a new image–text interaction network to investigate the relationship between affective image regions and text words for image–text sentiment analysis.

CGAFT [25] introduces an adaptive fine-tuning method to vision Transformer and employs Cross-Attention to fuse the patch-level image features and word-level text features. Therefore, it is a fine-grained patch-word interaction method.

1 https://wukong-dataset.github.io/wukong-dataset/
2 https://www.weibo.com
3 https://github.com/hkxiaodong/MCFIT/tree/main/CISD

Table 2
Experiment results of Acc and F1 on four datasets.

Model	MVSA-S		MVSA-M		TumEmo		CISD	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
$MMBT_{2019}$ [55]	73.17	72.43	72.21	70.63	65.36	65.05	65.36	65.05
$MMBT\text{-}F_{2020}$ [9]	73.39	73.14	72.39	69.86	64.48	64.14	64.48	64.14
$D\text{-}Senti_{2018}$ [8]	70.07	69.14	72.03	69.20	63.75	63.38	63.75	63.38
$DMAF_{2019}$ [14]	70.29	70.03	68.86	66.76	63.09	63.51	65.79	66.13
$Co\text{-}MN_{2018}$ [11]	70.51	70.01	69.92	69.83	64.26	59.09	–	–
$MVAN_{2021}$ [12]	72.98	72.98	72.36	72.30	66.46	63.39	–	–
$MGNNS_{2021}$ [13]	73.77	72.70	72.49	69.34	66.72	66.69	–	–
$TSAFNet_{2022}$ [34]	74.28	73.19	72.91	70.95	–	–	–	–
$CLMLF_{2022}$ [33]	75.33	73.46	72.00	69.83	–	–	–	–
$CGAFT_{2022}$ [25]	74.77	74.56	72.98	71.96	67.75	67.81	70.92	70.86
$ITIN_{2022}$ [19]	75.19	74.97	73.52	**73.49**	–	–	–	–
Region-word	75.61	75.43	74.32	73.27	65.05	65.38	69.29	69.38
Patch-word	76.94	76.59	74.44	71.38	69.78	69.90	75.27	75.34
MCFIT	**78.27**	**78.06**	**75.15**	72.47	**70.85**	**70.79**	**76.10**	**76.10**

Table 3
Ablation studies on four datasets.

Model	MVSA-S		MVSA-M		TumEmo		CISD	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
w/o Region	76.94	76.59	74.44	71.38	69.78	69.90	75.27	75.34
w/o Patch	75.61	75.43	74.32	**73.27**	65.05	65.38	69.29	69.38
w/o MWSA	78.05	77.49	74.97	70.58	70.18	70.23	74.62	74.79
w/o MECA	74.72	74.67	74.15	70.12	62.96	62.64	74.97	75.05
w/o MFN	76.94	76.62	74.56	72.00	70.61	70.60	75.61	75.74
w/o Memory	76.27	75.57	74.27	67.84	70.18	69.93	75.58	75.78
MCFIT	**78.27**	**78.06**	**75.15**	72.47	**70.85**	**70.79**	**76.10**	**76.10**

As shown in Fig. 1, to show that the collaborative fine-grained interaction we designed are effective, Cross-Attention is employed to implement two fine-grained region-word and patch-word interactions, which are denoted as **Region-word** and **Patch-word**. Note that, while **Region-word** and **Patch-word** are similar to the implementations in **INIT** and **CGAFT** respectively, they are stacked to multiple layers like the proposed MCFIT.

4.4. Comparison with the baselines

The performance comparisons between the proposed model and baselines are evaluated by accuracy and F1-score on four datasets. For convenience, the results in Table 2 are divided into three groups. The baselines in the first group are not fine-grained interaction methods and the baselines in second group are fine-grained interaction methods. Besides, MCFIT denotes our proposed model.

From all results in Table 2, we can get the following observations. **First.** Compared to the methods in the first group, the fine-grained interaction methods in the second group achieve better performance, which indicates that fine-grained interactions are beneficial for sentiment analysis. Besides, since both affective and non-affective patches are employed, patch-word interaction is more adaptable than region-word interaction. Thus, Patch-word obtains better results compared to Region-word. **Second.** The best performance of the proposed model is obtained due to the fact that both fine-grained interactions are employed and that we utilize patch and region information to improve region-word interaction and patch-word interaction learning, respectively.

4.5. Ablation studies

Several ablation experiments are conducted to quantify the contribution of each design in our proposed MCFIT model. **w/o Region.** We remove region branch and use Cross-Attention to learn patch-word interaction as shown in Fig. 1 (A.2). **w/o Patch.** Patch branch is removed and we employ Cross-Attention to model region-word interaction as shown in Fig. 1 (A.1). **w/o Memory.** Region-word and patch-word interactions are both modeled. **w/o MFN.** Concatenation is employed to replace Memory-augmented Fusion Network (MFN). Besides, Memory-enhanced Cross-Attention (**w/o MECA**) and Multi-Way Self-Attention (**w/o MWSA**) are also removed, respectively. Note that, since the implementation of MFN relies on memory, when we remove memory (i.e., w/o Memory), MFN is removed. Similarly, the removal of MECA is accompanied by the removal of memory, so the MFN is also removed. Finally, for fair comparison, the hyper-parameter settings (as shown in Section 4.8) in above ablation baselines are set to optimal values, respectively.

From the experimental results reported in Table 3 we can make the following conclusions. **(i)** When region-level (w/o Region) or patch-level (w/o Patch) features are removed, the performance of the model declines compared to MCFIT, which demonstrates that employing both fine-grained interactions facilitates the acquisition of more accurate sentiment predictions. **(ii)** The results of w/o Memory illustrates that the utilization of region and patch information improves patch-word interaction and region-word interaction learning, respectively. **(iii)** The superiority of MCFIT over w/o MFN proves that MFN can further utilize the information contained in memory vectors. **(iiii)** w/o MECA leads to suboptimal performance indicating that collaborative fine-grained interaction are more conducive to sentiment analysis. Moreover, the suboptimal results of w/o MWSA illustrate that intra-level relationships of three kinds of features further promote interaction learning.

Therefore, MCFIT consisting all sub-modules achieves the best performance and the removal of any one module would lead to suboptimal results.

4.6. Comparison with different memory styles

To leverage region and patch information to help patch-word interaction and region-word interaction learning respectively, memory vectors are introduced to implement MECA. Thus, to

Table 4
Comparison with different memory styles on four datasets.

Model	MVSA-S		MVSA-M		TumEmo		CISD	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Learnable [47]	76.94	76.62	73.80	**72.63**	70.10	70.29	74.90	74.91
Variant 1	77.38	76.58	75.05	72.09	70.50	70.66	75.35	75.36
Variant 2	77.16	76.70	74.38	66.80	70.61	70.38	75.50	75.51
Variant 3	78.05	77.84	74.97	72.40	70.54	70.56	75.95	76.00
MECA	**78.27**	**77.94**	**75.15**	72.47	**70.85**	**70.79**	**76.10**	**76.10**

Table 5
Collaborative fine-grained interaction comparison on four datasets.

Model	MVSA-S		MVSA-M		TumEmo		CISD	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Patch-word	76.94	76.59	74.44	71.38	69.78	69.90	75.27	75.34
Patch-word + region	77.58	77.19	74.86	71.64	70.25	70.21	75.62	75.69
Region-word	75.61	75.43	74.32	**73.27**	65.05	65.38	69.29	69.38
Region-word + patch	76.62	76.67	74.55	73.20	66.46	65.98	72.60	72.45
MCFIT	**78.27**	**78.06**	**75.15**	72.47	**70.85**	**70.79**	**76.10**	**76.10**

further verify the effectiveness of MECA, we compare it with several different memory styles. **Learnable.** Like [47], all memory vectors are learnable. **Variant 1.** As show in Fig. 3, m_v and m_k are updated by N_1. **Variant 2.** m_v and m_k are updated by N_2. **Variant 3.** Different from MECA, m_k is updated by N_1 and m_v is updated by N_2. The experimental results are listed in Table 4. Note that since all four memory styles have two hyperparameters, following Section 4.8, we set both hyperparameters to their optimal values for a fair comparison with MECA.

Compared to Learnable style, memory vectors of Variant 1 and Variant 2, which are dynamically updated, can provide information from image or text. Besides, since m_v and m_k are updated by different features, memory vectors of Variant 3 can provide more information. Compared with Variant 3, MECA gets better performance, which shows that the design of MECA is the most reasonable.

As a result, MECA is superior to other memory styles for achieving accurate collaborative fine-grained interaction.

4.7. Collaborative fine-grained interaction analysis

As shown in Fig. 1 (A.3 and B.3), our proposed collaborative fine-grained interaction realizes patch-word and region-word interactions at the same time, where patch and word information are utilized to improve region-word interaction and patch-word interaction learning, respectively. In this situation, is it effective to use only patch information to improve region-word interaction? And is it effective to employ only region information to improve patch-word interaction?

In order to answer these two questions, we construct two reduced versions of the collaborative fine-grained interaction as follows. **Region-word + patch.** Patches are utilized to provide non-affective information through MECA when modeling region-word interaction. **Patch-word + region.** Similarly, regions with bounding box information are leveraged through MECA to help locate affective patches when modeling the patch-word interaction. Meanwhile, the results of **Patch-word** and **Region-word** are added to better analyze the impact of using patch and region information.

From the experimental results reported in Table 5, we can get some conclusions. **(i)** Patch-word + region and Region-word + patch obtain better results compared to Patch-word and Region-word, which indicates that region and patch information contribute to patch-word interaction and region-word interaction learning. **(ii)** The metrics of Patch-word is higher on the large-scale dataset (TumEmo and CISD) and Region-word + patch achieves a greater improvement by adding patch information, which reflects that patch information can express the complete sentiment information of images. In other words, patch information is more conducive to improve sentiment analysis performance on large-scale image–text datasets. **(iii)** MCFIT achieves the best results indicating that the proposed collaborative fine-grained interaction can learn more accurate interactions.

4.8. Hyperparameter analysis

In the proposed model, stacking layer $\boldsymbol{N}$ and the length of memory vectors $\boldsymbol{M}$ are considered as two key hyperparameters influencing the performance of sentiment analysis. Thus, we conduct experiments on different settings of $\boldsymbol{N}$ and $\boldsymbol{M}$ to analyze the performance of MCFIT.

Hyperparameter analysis of $\boldsymbol{N}$. As shown in Table 1, the distributions of the datasets are relatively different. Thus, we conduct experiments on them under different N. Due to the limitation of computing resources, our experiments can only reach $N = 5$.

The results illustrated in Fig. 6 show that when $N = 5$ the accuracy and F1-score reach their maximum values on MSVA-S and MSVA-M. From the trends of F1-score line on MVSA-S and MVSA-M we can see that stacked encoder can improve the recognition rate of small sample sentiment categories in the unbalanced dataset. This also implies that deeper encoder is more conducive to obtaining accurate fine-grained interaction. The best performance is achieved with $N = 3$ on TumEmo and CISD indicating that large-scale and balanced dataset is more convenient to get accurate interaction. Thus, N is set to 5 and 3 in our experiments on (MVSA-S, MVSA-M) and (TumEmo, CISD), respectively.

Hyperparameter analysis of $\boldsymbol{M}$. We also report the experiments under different settings of memory vectors (m_k and $m_v \in \mathbb{R}^{M\times d}$). The results illustrated in Fig. 7 demonstrate that when $M = 30$ the accuracy and F1-score reach their maximum values on MVSA-S, MVSA-M, and TumEmo. Unlike them, $M = 20$ is the optimal setting on CISD.

Generally, proper setting of $\boldsymbol{N}$ and $\boldsymbol{M}$ will enhance collaborative fine-grained interaction learning, which improves the performance of image–text sentiment analysis.

4.9. Visualization

Attention Visualization. To better understand the interactions between image regions and words as well as image patches and

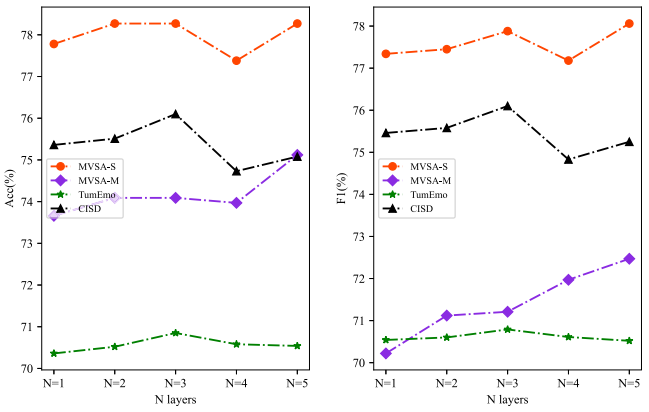

Fig. 6. The performance of different stacking layers.

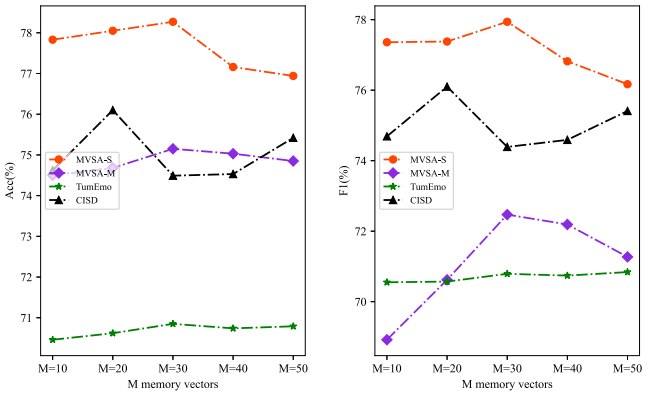

Fig. 7. The performance of different memory vectors.

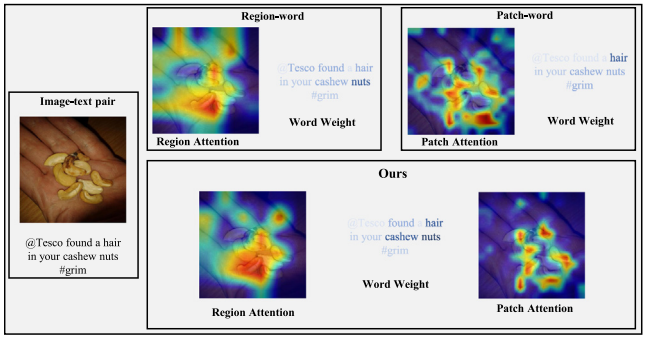

Fig. 8. Compared with Region-word and Patch-word, ours focuses better on the affective image regions and the corresponding words, which demonstrates that the proposed collaborative fine-grained interaction can yield more accurate image–text interaction. This also proves the validity of the proposed model.

words, we employ Grad-CAM [56] to visualize the attentions of image–text pairs learned by the models. As shown in Fig. 8, the attentions of an image are presented in heatmaps, and the corresponding text words with different learned weights are marked in different shades of color (the bigger the weight of the word, the heavier its color).

As illustrated in Fig. 8, compared with Patch-word, region information employed in our method can help locate the affective patches, which makes patch-word interaction more accurate. At the same time, compared with Region-word, the use of patch information provides auxiliary information for region-word interaction, which improves the region-word interaction learning. Therefore, the fine-grained interaction between image and text is

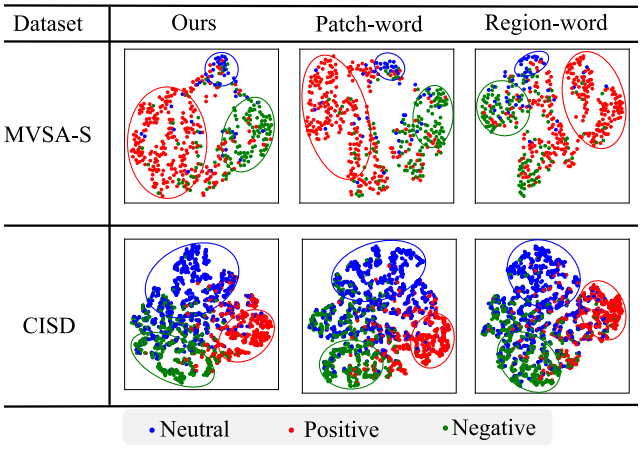

Fig. 9. Compared with Region-word and Patch-word, ours obtains better clusters in the feature spaces, which indicates that the proposed collaborative fine-grained interaction can better distinguish the sentiment information embedded in the image–text pairs.

better in the proposed method, which improves the performance of image–text interaction.

Cluster Visualization. To verify that the proposed collaborative fine-grained interaction can learn more accurate sentiment information in image–text data, cluster visualization is conducted on MVSA-S and CISD. Because of the higher performance on MVSA-S and CISD, their cluster visualizations are more representative. Specifically, t-SNE is employed to reduce the output of the models to 2-dimensional feature vector and visualize it.

As shown in Fig. 9, compared with Patch-word and Region-word, the features extracted by our method have better clusters in the feature space, which demonstrates that the proposed model can better distinguish the image–text sentiments embedded in image–text pairs.

5. Conclusion

In this paper, we propose a novel collaborative fine-grained interaction that employs region information and patch information to improve patch-word interaction and region-word interaction learning, respectively. Besides, to analyze the sentiments embedded in real-life Chinese image–text pairs, a large-scale Chinese image–text sentiment dataset named CISD is built. Finally, extensive experiments are conducted and the results demonstrate the effectiveness of the proposed method.

While very encouraging performance has been achieved, there are some limitations of our model. For example, since our model needs first encode image to region-level features via Faster R-CNN pre-trained on Visual Genomes dataset, it will be a disaster when the distribution of social images is far from Visual Genomes. This means that we cannot get proper region-level features, which would make our model perform sub-optimally. Besides, although our model is a 3-input network, the inputs of region and patch branches need to be extracted from the same modality (e.g., image), which limits the application and performance of our model in other 3-input tasks. Therefore, in our future work, we will try to develop an end-to-end model that can be widely applied to different multimodal tasks, including image–text and visual–text–audio sentiment analysis.

CRediT authorship contribution statement

Xingwang Xiao: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization.

Yuanyuan Pu: Supervision, Funding acquisition. **Dongming Zhou:** Writing – review & editing, Supervision. **Jinde Cao:** Writing – review & editing, Supervision. **Jinjing Gu:** Writing – review & editing, Supervision. **Zhengpeng Zhao:** Writing – review & editing, Supervision. **Dan Xu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by Foundation of China under Grant 61271361, 62162068, 61761046 and 62362070; in part by the Major Science and Technology Special Project in Yunnan Province under Grant 202302AF080006; in part by the Key Project of Applied Basic Research Program of Yunnan Provincial Department of Science and Technology under Grant 202001BB050043.

References

[1] L. Yue, W. Chen, X. Li, W. Zuo, M. Yin, A survey of sentiment analysis in social media, Knowl. Inf. Syst. 60 (2) (2019) 617–663.
[2] R. Kaur, S. Kautish, Multimodal sentiment analysis: A survey and comparison, Int. J. Serv. Sci. Manag. Eng. Technol. 10 (2) (2019) 38–58.
[3] R. Chen, W. Zhou, Y. Li, H. Zhou, Video-based cross-modal auxiliary network for multimodal sentiment analysis, IEEE Trans. Circuits Syst. Video Technol. (2022) 1.
[4] W. Guo, Y. Zhang, X. Cai, L. Meng, J. Yang, X. Yuan, LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition, IEEE Trans. Multimed. 23 (2020) 1785–1798.
[5] Z. Zeng, N. Xu, W. Mao, Event-driven network for cross-modal retrieval, in: Int. Conf. Inf. Knowledge Manage, 2020, pp. 2297–2300.
[6] S. Yang, Q. Li, W. Li, X. Li, A.A. Liu, Dual-level representation enhancement on characteristic and context for image-text retrieval, IEEE Trans. Circuits Syst. Video Technol. (2022) 1.
[7] N. Xu, Analyzing multimodal public sentiment based on hierarchical semantic attentional network, in: IEEE Int. Conf. Intell. Secur. Inform.: Secur. Big Data, ISI, 2017, pp. 152–154.
[8] A. Hu, S. Flaxman, Multimodal sentiment analysis to explore the structure of emotions, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2018, pp. 350–358.
[9] P. Basu, S. Tiwari, J. Mohanty, S. Karmakar, Multimodal sentiment analysis of # MeToo tweets using focal loss (grand challenge), in: BigMM, 2020, pp. 461–465.
[10] S. Thuseethan, S. Janarthan, S. Rajasegarar, P. Kumari, J. Yearwood, Multimodal deep learning framework for sentiment analysis from text-image web data, in: Proc. - IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol., WI-IAT, 2020, pp. 267–274.
[11] N. Xu, W. Mao, G. Chen, A co-memory network for multimodal sentiment analysis, in: Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., SIGIR, 2018, pp. 929–932.
[12] X. Yang, S. Feng, D. Wang, Y. Zhang, Image-text multimodal emotion classification via multi-view attentional network, IEEE Trans. Multimed. 23 (2021) 4014–4026.
[13] X. Yang, S. Feng, Y. Zhang, D. Wang, Multimodal sentiment detection based on multi-channel graph neural networks, in: ACL - Annu. Meet. Assoc. Comput. Linguist., Proc. Conf., 2021, pp. 328–339.
[14] F. Huang, X. Zhang, Z. Zhao, J. Xu, Z. Li, Image–text sentiment analysis via deep multimodal attentive fusion, Knowl.-Based Syst. 167 (2019) 26–37.
[15] J. Zhang, X. Liu, Z. Wang, H. Yang, Graph-based object semantic refinement for visual emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 32 (5) (2022) 3036–3049.
[16] G. Chen, J. Peng, W. Zhang, K. Huang, F. Cheng, H. Yuan, Y. Huang, A region group adaptive attention model for subtle expression recognition, IEEE Trans. Affect. Comput. (2021).
[17] S. Parthasarathy, C. Busso, Predicting emotionally salient regions using qualitative agreement of deep neural network regressors, IEEE Trans. Affect. Comput. 12 (2) (2018) 402–416.
[18] J.E. Steephen, S.C. Obbineni, S. Kummetha, R.S. Bapi, HED-ID: An affective adaptation model explaining the intensity-duration relationship of emotion, IEEE Trans. Affect. Comput. 11 (4) (2018) 736–750.
[19] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, J. Qian, Multimodal sentiment analysis with image-text interaction network, IEEE Trans. Multimed. (2022) 1.
[20] T. Zhou, J. Cao, X. Zhu, B. Liu, S. Li, Visual-textual sentiment analysis enhanced by hierarchical cross-modality interaction, IEEE Syst. J. 15 (3) (2021) 4303–4314.
[21] A. Pandey, D.K. Vishwakarma, VABDC-Net: A framework for visual-caption sentiment recognition via spatio-depth visual attention and bi-directional caption processing, Knowl.-Based Syst. 269 (2023) 110515.
[22] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.
[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, in: Int. Conf. Learning Representations, 2021.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[25] X. Xiao, Y. Pu, Z. Zhao, R. Nie, D. Xu, W. Qian, H. Wu, Image–text sentiment analysis via context guided adaptive fine-tuning transformer, Neural Process. Lett. (2022) 1–23.
[26] T. Niu, S. Zhu, L. Pang, A.E. Saddik, Sentiment Analysis on Multi-View Social Data, in: Lect. Notes Comput. Sci., Springer, 2016, pp. 15–27.
[27] V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, Utterance-level multimodal sentiment analysis, in: ACL - Annu. Meet. Assoc. Comput. Linguist., Proc. Conf., 2013, pp. 973–982.
[28] Q. You, J. Luo, H. Jin, J. Yang, Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia, in: WSDM - ACM Int. Conf. Web Search Data Min., 2016, pp. 13–22.
[29] N. Xu, W. Mao, Multisentinet: A deep semantic network for multimodal sentiment analysis, in: Int. Conf. Inf. Knowledge Manage, 2017, pp. 2399–2402.
[30] W. Liao, B. Zeng, J. Liu, P. Wei, J. Fang, Image-text interaction graph neural network for image-text sentiment analysis, Appl. Intell. (2022) 1–15.
[31] Y. Zhang, Y. Zhang, W. Guo, X. Cai, X. Yuan, Learning disentangled representation for multimodal cross-domain sentiment analysis, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–11.
[32] J. Yu, K. Chen, R. Xia, Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2022) 1.
[33] Z. Li, B. Xu, C. Zhu, T. Zhao, CLMLF:A contrastive learning and multi-layer fusion method for multimodal sentiment detection, in: Conf. North American Chapter Assoc. Comput. Linguist., NAACL, 2022.
[34] X. Hu, M. Yamamura, Two-stage attention-based fusion neural network for image-text sentiment classification, in: ACM Int. Conf. Proc. Ser., 2022, pp. 1–7.
[35] H. Zhu, Z. Zheng, M. Soleymani, R. Nevatia, Self-supervised learning for sentiment analysis via image-text matching, in: ICASSP IEEE Int. Conf. Acoust. Speech Signal Process Proc., 2022, pp. 1710–1714.
[36] Y. Liu, Z. Li, K. Zhou, L. Zhang, L. Li, P. Tian, S. Shen, Scanning, attention, and reasoning multimodal content for sentiment analysis, Knowl.-Based Syst. 268 (2023) 110467.
[37] Y. Du, Y. Liu, Z. Peng, X. Jin, Gated attention fusion network for multimodal sentiment classification, Knowl.-Based Syst. 240 (2022) 108107.
[38] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, R. Xu, Multi-modal sarcasm detection via cross-modal graph convolutional network, in: ACL - Annu. Meet. Assoc. Comput. Linguist., Proc. Conf., 2022, pp. 1767–1777.
[39] R. Ratcliff, A theory of memory retrieval, Psychol. Rev. 85 (2) (1978) 59.
[40] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.
[41] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[42] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.
[43] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, R. Salakhutdinov, Transformer-XL: Attentive language models beyond a fixed-length context, in: ACL - Annu. Meet. Assoc. Comput. Linguist., Proc. Conf., 2019, pp. 2978–2988.
[44] A. Al Adel, M.S. Burtsev, Memory transformer with hierarchical attention for long document processing, in: Int. Conf. Eng. Telecommun., En T, 2021, pp. 1–7.
[45] J. Lei, L. Wang, Y. Shen, D. Yu, T. Berg, M. Bansal, MART: Memory-augmented recurrent transformer for coherent video paragraph captioning, in: ACL - Annu. Meet. Assoc. Comput. Linguist., Proc. Conf., 2020, pp. 2603–2614.
[46] L.V. Cardoso, S.J.F. Guimaraes, Z.K. Patrocínio, Enhanced-memory transformer for coherent paragraph video captioning, in: Proc. Int. Conf. Tools Artif. Intell., ICTAI, 2021, pp. 836–840.

[47] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit, 2020, pp. 10578–10587.
[48] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit, 2018, pp. 6077–6086.
[49] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit, 2016, pp. 770–778.
[51] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: Int. Conf. Machine Learning, 2021, pp. 8748–8763.
[52] K. Wu, H. Peng, M. Chen, J. Fu, H. Chao, Rethinking and improving relative position encoding for vision transformer, in: Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit, 2021, pp. 10033–10041.
[53] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, in: Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit, 2018, pp. 3588–3597.
[54] J. Gu, X. Meng, G. Lu, L. Hou, M. Niu, X. Liang, L. Yao, R. Huang, W. Zhang, X. Jiang, Wukong: 100 million large-scale Chinese cross-modal pre-training dataset and a foundation framework, 2022, arXiv.
[55] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, D. Testuggine, Supervised multimodal bitransformers for classifying images and text, 2019, arXiv preprint arXiv:1909.02950.
[56] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proc. IEEE Int. Conf. Comput. Vision, 2017, pp. 618–626.