BIT: Improving Image-text Sentiment Analysis via Learning Bidirectional Image-text Interaction

Xingwang Xiao, Yuanyuan Pu✉, Zhengpeng Zhao, Jinjing Gu, Dan Xu
School of Information Science and Engineering
Yunnan University
Kunming 650504, China
576950475@qq.com, {yuanyuanpu, zhpzhao, jinjinggu, danxu}@ynu.edu.cn

***Abstract*—Exploring the interaction between image and text has a great strength for image-text sentiment analysis. However, most methods only focus on learning forward interaction in forward image-text features and fail to capture the backward interaction in backward image-text features, which leads to the loss of necessary information embedded in backward interaction. In this paper, Bidirectional Interaction Transformer (BIT) that models both forward and backward image-text interactions is proposed for image-text sentiment analysis. Specifically, we first encode image and text to forward and backward features. Then, these features are fed into Bidirectional Interaction Encoder (BIE) with Forward Interaction and Back Interaction branches to model bidirectional (i.e., forward and backward) image-text interaction. Finally, Two-scale Adaptive Gating Fusion (TAGF) is designed to adaptively fuse the forward and backward interactions learned by BIE. Extensive experiments conducted on two public datasets demonstrate the effectiveness of the proposed model.**

***Index Terms*—Image-text sentiment analysis, Bidirectional image-text interaction, Bidirectional transformer.**

I. Introduction

With the increased use of mobile terminal devices and the bloom of social media platforms such as Twitter, Tumblr, and Weibo, more and more users post multimodal tweets (e.g., text, image, and video) about diverse events and topics to express their feelings and emotions. Therefore, in recent years, research with the purpose of analyzing the sentiments embedded in multimodal data has received increasing attention. More specifically, exploring multimodal sentiments has many applications, including multimodal sentiment analysis [1], decision making [2], cross-modal information retrieval [3], multimodal fake news detection [4] and so on.

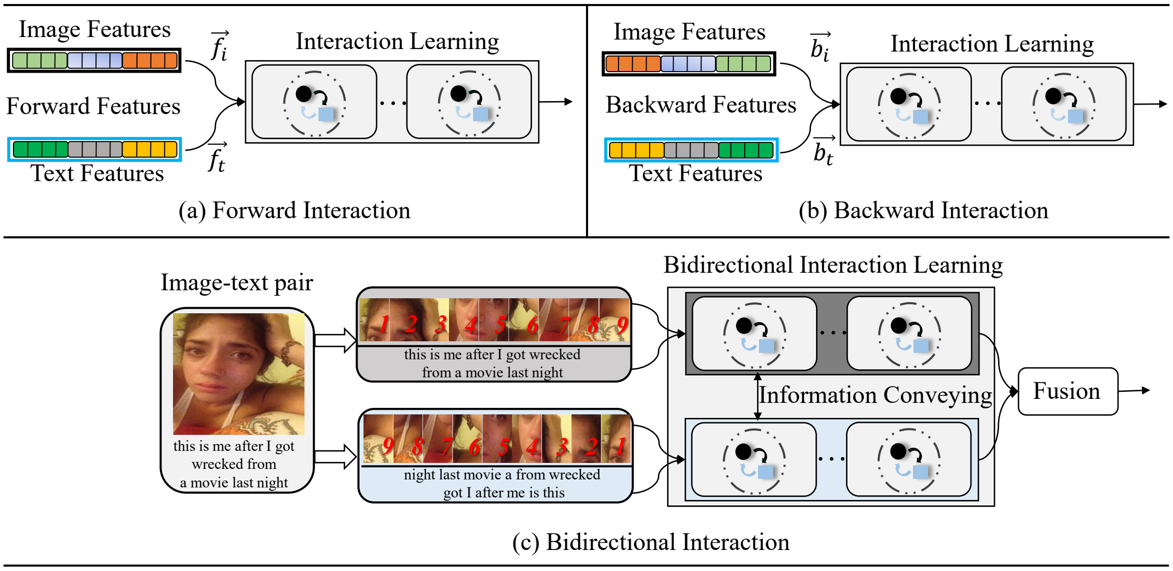
For multimodal sentiment analysis, we focus on image-text sentiment analysis in social media data. In existing methods, some works fuse different modality features via simple concatenation [5]–[8], which could not effectively explore the relationship between image and text. Meanwhile, some studies focus on modeling the relationship and interaction of image-text pair [9]–[12]. In general, the affective regions of an image are able to evoke the sentiments of human [13], [14]. Thus, some recent works have tried to model the image-text interaction between image regions and text words [15], [16]. However, these methods only model the forward interaction between the forward features as shown in Fig 1 (a).

Since the studies including BRNN [17], Bi-LSTM [17], [18], and Bi-GRU [17], [19] have shown that bidirectional (i.e., forward and backward) information learned from forward and backward inputs can improve the performance of models, Bi-LSTM and Bi-GRU have been employed to learn the bidirectional information from unimodal data (i.e., image or text) for image-text sentiment analysis [6], [11], [15]. Motivated by this, we believe that learning both the forward interaction information in the forward features (Fig 1 (a)) and the backward interaction information in the backward features (Fig 1 (b)) can enhance the diversity of interaction information, which will benefit image-text sentiment analysis. However, while there are methods that utilize bidirectional information learned from unimodal data, no studies have explored bidirectional information at image-text interaction level and most existing studies only focus on learning the forward image-text interaction, which may result in the loss of necessary information embedded in backward image-text interaction.

Fortunately, since an image is worth 16×16 words (patches) [20], we can consider the image-text pair as forward and backward patch-word sequences, which allows us to learn bidirectional (i.e., forward and backward) image-text interaction information from bidirectional sequences. Ideally, as shown in Fig 1 (c), we expect the proposed model is able to learn bidirectional image-text interaction and convey interaction information between two branches. Meanwhile, we hope that the conveyed interaction information will enhance the forward and backward image-text interaction learning.

To realize the above goals, in this paper, we propose Bidirectional Interaction Transformer (BIT) for image-text sentiment analysis, which models both forward and backward image-text interactions and achieves information conveying. The proposed model can be divided into three stages. **(i) Feature Encoding.** Given an image-text pair, the image is encoded to forward and backward patch-level features, and the text is encoded to forward and backward word-level features. **(ii) Bidirectional Image-text Interaction Learning.** We devise a Bidirectional Interaction Encoder (BIE) that consists of Forward Interaction (FI) and Backward Interaction (BI) branches to model forward image-text interaction between forward features and learn

✉Corresponding author: yuanyuanpu@ynu.edu.cn



(a) Forward Interaction

(b) Backward Interaction

(c) Bidirectional Interaction

Fig. 1. Three types of image-text interactions. Among them, (a) and (b) learn forward and backward image-text interactions, respectively. In contrast to (a) and (b), the bidirectional interaction we propose in (c) models both forward and backward interactions and conveys information between the two branches. As shown in (c), given an image-text pair, two branches are employed to solve the problem of how to learn forward image-text interaction and backward image-text interaction respectively. Meanwhile, we also address the issue of how to convey interaction information between two branches to enhance forward and backward image-text interaction learning.

backward image-text interaction between backward features respectively. Furthermore, since forward and backward interaction information can be conveyed by Prefix Interaction Network (PIN), Prefix-enhanced Interaction Attention (PIA) of BI and FI enhanced by this information will improve backward and forward interaction learning. Meanwhile, BIE is stacked to N layers to further explore the deep bidirectional interaction. **(iii) Adaptive Fusion.** To adaptively fuse the forward and backward image-text interactions learned by N-layer BIE, we design Two-scale Adaptive Gating Fusion (TAGF).

The main contributions of this paper can be summarized as follows.

- We propose a novel bidirectional interaction network BIT that can model bidirectional image-text interaction between image and text to achieve image-text sentiment analysis.
- We utilize bidirectional information at image-text interaction level to improve image-text sentiment analysis. Meanwhile, we also achieve interaction information conveying in our encoder to enhance the forward and backward image-text interaction learning, which further improves the performance of sentiment analysis.
- Extensive experiments are conducted on two public datasets, and the results demonstrate that our model outperforms the state-of-the-art methods. Ablation and bidirectional interaction studies are also conducted and the results illustrate the effectiveness of the proposed method.

For convenience, in the remainder of this paper, "image-text interaction" is also shortened to "interaction". For example, bidirectional interaction denotes bidirectional image-text interaction.

II. Related work

A. *Image-text Sentiment Analysis*

Image-text sentiment analysis exploits the complementary of information between image and text to make more accurate sentiment prediction.

Recently, with the development of deep learning, studies based on deep learning have achieved surprising performance. Xu et al. developed different models including HSAN [7], MultiSentiNet [8], and Co-Memory [9] for image-text sentiment analysis. MVAN [10] and MGNNS [11] were also successively proposed by Yang et al. for image-text sentiment analysis. Zhang et al. [21] proposed a disentangled sentiment representation adversarial network to reduce the domain shift of expressive styles for image-text cross-domain sentiment analysis. To achieve aspect-based image-text sentiment analysis, Yu et al. [22] developed a hierarchical interactive multimodal transformer. Two contrastive learning tasks were adopted by [23] to construct a contrastive learning and multi-layer fusion model. Based on attention mechanism, Hu et al. [12] proposed a two-stage attention-based fusion neural network for image-text sentiment analysis. Image-text matching with self-supervised learning was utilized by [24] to achieve image-text sentiment analysis. On insufficient labeled data, intermediate fusion and late fusion were adopted by Kumar et al. [25] to propose a hybrid fusion based method for image-text emotion recognition. On imbalance data, Basu et al. [5] constructed a multimodal network for Metoo tweet sentiment analysis. Wang et al. [26] developed a multimodal event-aware network for image-text sentiment analysis in tourism.

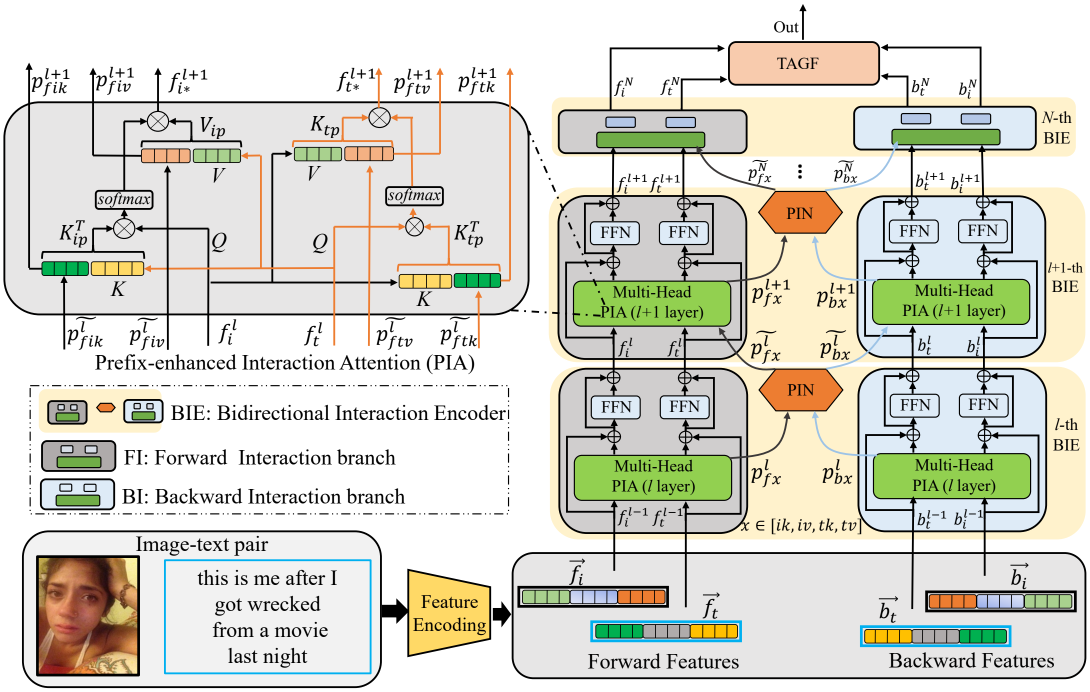


Fig. 2. The framework of Bidirectional Interaction Transformer. For an image-text pair, it is encoded to forward and backward features. Then, these features will be fed into N-layer BIE to model bidirectional interaction, where FI and BI model forward interaction and backward interaction respectively. PIN (Fig 3) is employed to convey interaction information between FI and BI,which enhances the sentimental image-text interaction learning of PIA. In particular, for the N-th BIE, PIN is removed. Finally, we design TAGF (Fig 4) to fuse forward and backward interaction information adaptively.

Hu et al. [6] concatenated image and text features to achieve image-text emotion recognition. Jiang et al. [27] constructed a fusion-extraction network for image-text sentiment analysis. Zhu et al. [15] proposed an image-text interaction network to model the relationship between image regions and text words. Wang et al. [28] developed an end2end fusion method with transformer for image-text sentiment analysis. A sentiment multi-layer neural network was developed by [29], which combined several visual features with contextual text features to predict the overall sentiment. Huang et al. [30] proposed a deep multimodal attentive fusion model to achieve image-text sentiment analysis. Yadav et al. [31] proposed a deep multi-level attentive network for image-text sentiment analysis.

Although above image-text sentiment analysis methods have achieved promising results, most of them only focus on modeling forward interaction of image-text pair and fail to capture the backward interaction of image-text. In this paper, we propose BIT that can model bidirectional interaction between image and text for image-text sentiment analysis.

B. Bidirectional Networks

Bidirectional networks aim to obtain information from different directions. Thus, many studies about learning bidirectional information have been applied to artificial neural networks. For example, some works including BRNN [17], Bi-LSTM [17], [18], and Bi-GRU [17], [19] have proved that bidirectional information can improve the performance of the models.

More recently, Xu et al. [32] proposed a bidirectional prediction method by using multi-generative multi-adversarial nets. For fake news detection, a bidirectional cross-modal fusion model was developed by Yu et al. [33]. Devlin et al. [34] pre-trained deep bidirectional transformers for language understanding. Bidirectional backpropagation was employed by [35] for bidirectional associative memory. Wang et al. [36] proposed a lightweight bidirectional feedback network for image super-resolution. Zhang et al. [37] leveraged bidirectional encoding structure with channel attention cascade for video frame interpolation.

From above methods we find that the learned bidirectional information can further improve the performance of the models. Motivated by this, a bidirectional interaction network BIT is proposed to model bidirectional interaction at image-text interaction level for image-text sentiment analysis

III. Methodology

In this section, we elaborate on the details of the proposed BIT for multimodal sentiment analysis. The framework of BIT is shown in Fig 2.

A. Feature encoding

Forward and Backward Word-level Features. To represent a sentence $\boldsymbol{S}$ with k words, we employ Transformer stream of CLIP (ViT-B/16) [38] pre-trained on 400 million image-text pairs and remove the last features choosing operation to encode each sentence to word-level features. In this case, each word of a sentence is embedded to 512-dimensional word-level feature w_i $(i = 1, 2, \ldots, k)$. Therefore, we can get forward word-level features $\overrightarrow{f_t} = \{w_1, \ldots, w_k\}$.

To represent backward word-level features of the sentence, we reverse forward word-level features. Thus, we can get the backward text features $\overrightarrow{b_t} = \{w_k, \ldots, w_1\}$.

Forward and Backward Patch-level Features. As stated in Vision Transformer (ViT) [20], an image is worth 16×16 words. In other words, an image is worth 16×16 patches. Motivated by this, to represent an image $\boldsymbol{I}$ with n patches, we employ ViT stream of CLIP (ViT-B/16) [38] and remove the last features choosing operation to encode each image patch to 512-dimensional patch-level feature p_i $(i = 1, 2, \ldots, n)$. Thus, we can get forward patch-level features $\overrightarrow{f_i} = \{p_1, \ldots, p_n\}$.

Similarly, we also reverse forward patch-level features of image to obtain backward patch-level features. Therefore, we get the backward image features $\overrightarrow{b_i} = \{p_n, \ldots, p_1\}$.

In general, given an image-text pair $(\boldsymbol{I}, \boldsymbol{T})$, it is encoded to forward patch-word feature sets $(\overrightarrow{\boldsymbol{f_i}}, \overrightarrow{\boldsymbol{f_t}})$ and backward patch-word feature sets $(\overrightarrow{\boldsymbol{b_i}}, \overrightarrow{\boldsymbol{b_t}})$. Then, these features will be sent to N-layer encoder to learn bidirectional interaction, where forward interaction and backward interaction are learned from $(\overrightarrow{\boldsymbol{f_i}}, \overrightarrow{\boldsymbol{f_t}})$ and $(\overrightarrow{\boldsymbol{b_i}}, \overrightarrow{\boldsymbol{b_t}})$, respectively.

B. Bidirectional Interaction Encoder

We elaborately develop a stackable Bidirectional Interaction Encoder (BIE) to model the bidirectional interaction between image patches and text words. As shown in Fig 2, BIE consists of two parallel interaction branches with Prefix-enhanced Interactive Attention (PIA) (left and right denote Forward Interaction (FI) branch and Backward Interaction (BI) branch respectively) and a Prefix Interaction Network.

Their implementations are based on scaled dot-product Cross-Attention (CA) [39]. Specifically, CA operates on three sets of vectors, namely a set of queries Q, keys K and values V, and takes a weighted sum of value vectors according to a similarity distribution between query and key vectors. For convenience, CA is formally defined as:

$$\mathrm{CA}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \qquad (1)$$

where Q is a matrix of n_q query vectors, K and V contain n_k keys and values, all with the same dimensionality, and $\sqrt{d}$ is a scaling factor.

Prefix Interaction Network. To bridge the two parallel interaction branches (FI and BI) and convey information between them, we design a Prefix Interaction Network (PIN). Note that, the conveyed information can enhance the image-text interaction learning of PIA in FI and BI, which further improves the performance of sentiment analysis.

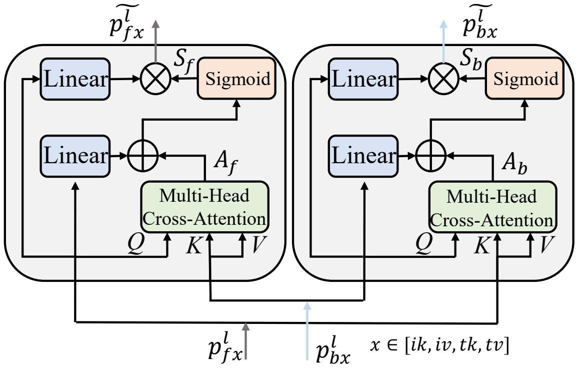


Fig. 3. The architecture of Prefix Interaction Network. PIN can convey interaction information between FI and BI, which can enhance the sentimental image-text interaction learning of PIA.

As visualized in Fig 3, assume that p^l_{fx} and p^l_{bx} $(x \in [ik, iv, tk, tv])$ are the prefixes from l-th FI and BI respectively, p^l_{fx} and p^l_{bx} will be fed into PIN to model the interaction between them. Since the sub-models of the left and right are the same, for convenience, the interaction process of the left is employed as paradigm. The interaction process is defined as follows:

$$\begin{aligned} A_f &= \mathrm{MHCA} = \mathrm{Concat}(\mathrm{head}_1, \cdots, \mathrm{head}_n)W_f, \\ &\text{where, } \mathrm{head}_i = \mathrm{CA}\left(W^i_q p^l_{fx}, W^i_k p^l_{bx}, W^i_v p^l_{bx}\right), \end{aligned} \qquad (2)$$

$$\begin{aligned} S_f &= \mathrm{Sigmoid}\left(A_f \oplus \left(w_1 p^l_{fx} + b_1\right)\right), \\ \widetilde{P^l_{fx}} &= S_f \otimes \left(w_2 p^l_{fx} + b_2\right), \end{aligned} \qquad (3)$$

where MHCA is Multi-Head Cross-Attention [39], W^i_q, W^i_k, W^i_v, w_1, w_2, and W_f are learnable weights, b_1 and b_2 are learnable biases, $\otimes$ and $\oplus$ denote element-wise product and element-wise addition, respectively.

As shown in Eq.(2), p^l_{fx} and p^l_{bx} are first sent into MHCA to get interaction A_f between them. A_f denotes that p^l_{fx} is used as query to obtain backward information from p^l_{bx}. Therefore, the backward sentiment and interaction information embedded in p^l_{bx} will be incorporated into A_f. Then, as defined in Eq.(3), A_f is added to p^l_{fx} before scaling by sigmoid. Thus, the update gate S_f can determine how much information from p^l_{fx} and A_f needs to be preserved. Clearly, the output prefixes ($\widetilde{P^l_{fx}}$ and $\widetilde{P^l_{bx}}$) that contain sentiment and interaction information from p^l_{fx} and p^l_{bx} will contribute to obtaining accurate sentimental interaction in $(l + 1)$-th FI and BI of stacked BIE.

Generally, PIN can convey forward interaction information and backward interaction information between FI and BI, which enhances the sentimental image-text interaction learning of PIA.

Prefix-enhanced Interaction Attention. Usually, CA is used to fuse information between different modalities or to obtain interaction information. However, the interaction information conveyed by PIN can not be directly merged by vanilla CA. Therefore, to model the sentiment interaction between forward features as well as backward features, and also to utilize the information conveyed between FI and BI, Prefix-enhanced Interaction Attention (PIA) is elaborately constructed by us and it is visualized in Fig 2. In PIA, the sets of keys and values used for CA (Eq.(1)) are extended with two additional prefix vectors that contain interaction information from FI and

BI. Note that, for the first BIE, the prefixes are initialized according to the input features. Formally, for the (l+1)-th encoder, PIA and Multi-Head PIA (MHPIA) is defined as:

$$\text{PIA} \;=\left\{\begin{array}{l} CA_i\left(Q, K_{ip}, V_{ip}\right)=\operatorname{softmax}\left(\frac{Q K_{ip}^T}{\sqrt{d}}\right) V_{ip} \\ CA_t\left(Q, K_{tp}, V_{tp}\right)=\operatorname{softmax}\left(\frac{Q K_{tp}^T}{\sqrt{d}}\right) V_{tp} \end{array}\right. ,$$
$$\text{where, } K_{ip}=[K, \widetilde{p_{fik}^l}], V_{ip}=[V, \widetilde{p_{fiv}^l}],$$
$$K_{tp}=[K, \widetilde{p_{ftk}^l}], V_{tp}=[V, \widetilde{p_{ftv}^l}] \tag{4}$$

$$\text{MHPIA}=\text{Concat}\left(\text{head}_1, \cdots, \text{head}_n\right) W_m, \;\text{where,}$$
$$\text{head}_i=\left\{\begin{array}{l} CA_i\left(W_{qi}^i f_i,[W_{ki}^i f_t, \widetilde{P_{fik}^l}],[W_{vi}^i f_t, \widetilde{P_{fiv}^l}]\right) \\ CA_t\left(W_{qt}^i f_t,[W_{kt}^i f_i, \widetilde{P_{ftk}^l}],[W_{vt}^i f_i, \widetilde{P_{ftv}^l}]\right) \end{array}\right. , \tag{5}$$

where $W_{qi}^i, W_{qt}^i, W_{ki}^i, W_{kt}^i, W_{vi}^i, \underset{\sim}{W_{vt}^i,}$ and W_m are learnable weights, [,] denotes concatenation. $\widetilde{P_{fx}^l}(x \in [ik, iv, tk, tv])$ is from $\underset{\sim}{\text{PIN}}$ and it is calculated following Eq.(2) and Eq.(3). Since $\widetilde{P_{fx}^l}$ contains backward sentiment interaction information from BI, the PIA of FI enhanced by $\widetilde{P_{fx}^l}$ will contribute to modeling more accurate forward sentiment interaction between forward features. On the other hand, the PIA of BI enhanced by forward interaction information will contribute to modeling more accurate backward sentiment interaction. Particularly, the PIN is removed as the last layer of N-layer BIE does not continue to convey information. Then, like Transformer [39], FFN (two linear layers) and residual connections are employed in FI and BI.

To sum up, BIE consists of two parallel interaction branches (FI and BI) and they have the same structure, where FI models forward sentiment interaction between forward features and BI models backward sentiment interaction between backward features. Accordingly, BIE can model bidirectional interaction between image and text. Besides, PIN that can convey forward and backward information between FI and BI further enhances the sentimental image-text interaction learning of PIA. Furthermore, we are able to further explore deep bidirectional interaction via N-layer BIE.

C. Two-scale Adaptive Gating Fusion

Since our encoder BIE consists of FI and BI branches, the final output contains forward and backward sentimental interaction information. In order to fully utilize forward and backward information and fuse forward interaction and backward interaction information adaptively, we thoughtfully design a Two-scale Adaptive Gating Fusion network (TAGF).

As demonstrated in Fig 4, the outputs of N-layer BIE include forward outputs (f_i^N and f_t^N) and backward outputs (b_i^N and b_t^N). In order to effectively retain the forward and backward sentiment interaction information, the outputs are first sent to Two-scale Concatenation (TC) for information compression and fusion. Then, fused forward information f_o and fused backward information b_o are fed into Adaptive Gating (AG) to get scalar λ $(0<\lambda<1)$, which indicates the importance of backward interaction and forward interaction. We summarize this procedure as:

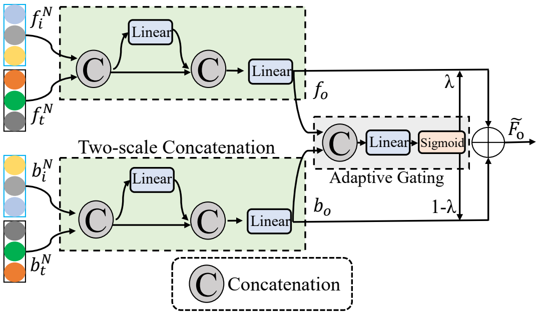


Fig. 4. The framework of Two-scale Adaptive Gating Fusion. Two-scale Concatenation aims to retain more information and the λ learned by Adaptive Gating indicates the improtance of f_o and b_o.

$$\begin{gathered} f_o=w_2\left(w_1\left(\text{Concat}\left(f_i^N, f_t^N\right)\right)+b_1\right)+b_2, \\ b_o=w_4\left(w_3\left(\text{Concat}\left(b_i^N, b_t^N\right)\right)+b_3\right)+b_4, \\ \lambda=\text{Sigmoid}\left(w_5\left(\text{Concat}\left(f_o, b_o\right)+b_5\right)\right., \\ \widetilde{F_o}=\lambda f_o+(1-\lambda) b_o, \end{gathered} \tag{6}$$

where w_1, w_2, w_3, w_4, and w_5 are learnable weights, b_1, b_2, b_3, b_4, b_5 are learnable biases. $\widetilde{F_o}$ is fused bidirecional interaction representation, which can further improve the performance of image-text sentiment analysis.

D. Image-text Sentiment Classification

The objective of BIT is to identify which sentiment or emotion is embedded in an image-text input. Hence, we feed the fused bidirectional sentiment representation $\widetilde{F_o}$ into softmax for image-text sentiment classification. And our model is optimized by minimizing the cross-entropy loss.

$$\begin{gathered} \hat{y}=\operatorname{softmax}\left(\widetilde{F_o}\right), \\ \mathcal{L}=-\sum_i y_i \log \hat{y}_i, \end{gathered} \tag{7}$$

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

TABLE I
DETAILED STATISTICS OF TWO DATASETS

	Positive	Neutral	Negative	All
MVSA-S	2,683	470	1,358	4,511
		TumEmo		
Angry			14,554	
Bored			32,283	
Calm			18,109	
Fear			20,264	
Happy			50,267	
Love			34,511	
Sad			25,277	
All			195,265	

IV. Experiments

A. Image-text Datasets

Two public datasets including MVSA-Single (MVSA-S) [40] and TumEmo [10] are employed to evaluate our proposed method BIT. Specifically, MVSA-S is an image-text sentiment dataset collected from Twitter, and TumEmo is a large-scale image-text emotion dataset collected from Tumblr.

For fair comparison, we preprocess MVSA-S following the method in [8] and employ the same method provided in [10] to preprocess TumEmo. Finally, the detailed statistics of two datasets used for our experiments are listed in Table I. In our experiments, the same training set, valid set, and test set of both datasets are employed according to [8], [10]

B. Implementation Details

We utilize Adam to optimize our model that is implemented by PyTorch. Batch sizes of MVSA-S and TumEmo are set to 16, 32 respectively. Head number is set to 8. The initial learning rate is 1e-4 and 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, step size and gama are 2 and 0.8 on TumEmo.

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

C. Image-text Sentiment Analysis Baselines

We compare our model with the following image-text sentiment analysis baselines. **MultiSentiNet [8]** extracts dual-view features of image for image-text sentiment analysis. **Co-Memory [9]** is a co-memory network that iteratively models the relations between image and text. **MMBT [41]** is reported by Facebook AI Research for image-text classification. **MMBT-F [5]** is based on MMBT and aims to achieve sentiment analysis of imbalanced image-text datasets. **MVAN [10]** is a multi-view attentional network that utilizes co-memory network for image-text emotion analysis. **MGNNS [11]** develops multi-channel graph neural networks to learn multimodal representations based on the global characteristics of the dataset. **TSAFNet [12]** first uses CA to explore correlation between image and text, and then a dual attention network is employed to analyze the features in channel and spatial dimension. **ITIN [15]** investigates the relationship between affective image regions and text words for image-text sentiment analysis. **CLMLF [23]** employs two contrastive learning tasks to help the model learn common features related to sentiment in image-text data. **CLIP-C** concatenates image-text features encoded by CLIP [38]. **CLIP-BiLSTM** learns bidirectional information from unimodal features via two Bi-LSTMs. **CLIP-FI** adopts FI with CA to model forward interaction between the image-text features encoded by CLIP. **CLIP-FI+BI** employs FI and BI with CA to model bidirectional information at image-text interaction level. For fair comparison, as in our model BIT, CLIP is used as feature extractor (i.e., parameters are frozen). Note that, the results of CLIP-FI and CLIP-FI+BI are the same as w/o BI and w/o PIN in Table III, respectively.

TABLE II
Experimental results of Acc and F1 on two datasets.

Model	MVSA-S		TumEmo	
	Acc	F1	Acc	F1
MultiSentiNet$_{2018}$ [8]	69.84	69.63	64.18	56.92
Co-Memory$_{2019}$ [9]	70.51	70.01	64.26	59.09
MMBT$_{2019}$ [41]	73.17	72.43	65.36	65.05
MMBT-F$_{2020}$ [5]	73.39	73.14	64.48	64.14
MVAN$_{2021}$ [10]	72.98	72.98	66.46	63.39
MGNNS$_{2021}$ [11]	73.77	72.70	66.72	66.69
TSAFNet$_{2022}$ [12]	74.28	73.19	67.54	67.10
ITIN$_{2022}$ [15]	75.19	74.97	–	–
CLMLF$_{2022}$ [23]	75.33	73.46	68.74	68.83
CLIP-C	76.99	76.78	68.66	68.86
CLIP-BiLSTM	77.18	77.06	69.65	69.44
CLIP-FI	77.38	77.23	70.74	70.88
CLIP-FI+BI	77.83	77.85	70.87	70.95
BIT	**79.16**	**78.91**	**71.62**	**71.68**

TABLE III
Abaltion studies on two datasets.

Model	MVSA-S		TumEmo	
	Acc	F1	Acc	F1
w/o BI	77.38	77.23	70.74	70.88
w/o FI	76.88	76.75	67.65	67.11
w/o PIN	77.83	77.85	70.87	70.95
w/o TAGF	77.16	77.14	70.31	70.02
w/o AG	78.05	78.39	71.06	71.03
w/o TC	77.61	76.84	70.87	70.94
BIT	**79.16**	**78.91**	**71.62**	**71.69**

D. Comparison with Baselines

The results of all models are shown in Table II, where the results in bold are optimal, and the underlined results are suboptimal.

From the reported results we can make the following observations. **First.** Since Bi-LSTM can learn bidirectional information from unimodal features, CLIP-BiLSTM achieves better results than CLIP-C. **Second.** Compared with CLIP-FI, the backward interaction information is utilized by CLIP-FI+BI to further improve the performance. **Third.** Since BIT can convey information between FI and BI, image-text interaction learning is further enhanced. Thus, BIT gets the best results.

E. Ablation Studies

We conduct several ablation studies on two datasets to quantify the contribution of each design in our proposed BIT. More specifically, we remove FI branch, BI branch, PIN, and TAGF respectively, which are denoted as w/o BI, w/o FI, w/o PIN, and w/o TAGF. Besides, to verify the effectiveness of the sub-modules in TAGF, we remove TC and AG respectively, which are denoted as w/o TC and w/o AG. Note that, since PIN aims to convey information between FI and BI, when FI or BI is removed, PIN is also removed. Besides, since PIA is designed to leverage the interaction information conveyed between FI and BI, when FI or BI or PIN is removed, we use CA instead of PIA.

From the results reported in Table III we can make the following observations. **(i)** The removal of any one module of

TABLE IV
RESULTS OF DIFFERENT N AND P.

Dataset	Acc	F1
MVSA-S	77.61$_{(P=10,N=1)}$	77.33$_{(P=10,N=1)}$
	78.05$_{(P=20,N=6)}$	77.52$_{(P=20,N=6)}$
	79.16$_{(P=30,N=3)}$	**78.91**$_{(P=30,N=3)}$
	77.83$_{(P=40,N=1)}$	77.48$_{(P=40,N=1)}$
	78.05$_{(P=50,N=6)}$	78.06$_{(P=50,N=6)}$
TumEmo	71.14$_{(P=10,N=6)}$	71.08$_{(P=10,N=6)}$
	71.43$_{(P=20,N=4)}$	71.32$_{(P=20,N=4)}$
	71.62$_{(P=30,N=5)}$	**71.69**$_{(P=30,N=5)}$
	71.53$_{(P=40,N=5)}$	71.62$_{(P=40,N=5)}$
	71.46$_{(P=50,N=2)}$	71.47$_{(P=50,N=2)}$

BIT would lead to suboptimal results of sentiment analysis. **(ii)** When FI or BI is removed, the performance of the model declines compared to BIT, which demonstrates that bidirectional sentiment interaction facilitates the acquisition of more accurate sentiment prediction. **(iii)** While the results of w/o FI are poor, they do illustrate that there is necessary information embedded in backward interaction. **(iv)** w/o PIN leads to suboptimal results indicating that interaction information conveying between FI and BI further promote image-text interaction learning of PIA. **(v)** The results of w/o TC and w/o AG outperform w/o TAGF show that TC can retain more information and AG can adaptively fuse forward information and backward information to enhance sentiment analysis.

F. Hyperparameter Analysis

In our experiments, stacking layer $\boldsymbol{N}$ and prefix length $\boldsymbol{P}$ are considered as key hyperparameters that influence the performance of sentiment analysis. Therefore, we conduct experiments under different settings of $\boldsymbol{N}$ and $\boldsymbol{P}$ to analyze the performance of BIT.

Results of different $\boldsymbol{N}$ and $\boldsymbol{P}$. To obtain the optimal performance of BIT, we conduct experiments under different N ($N \in [1,2,3,4,5,6]$) and P ($P \in [10,20,30,40,50]$). From the results listed in Table IV we can find that N and P influence differently on different datasets, where (P=x, N=y) denotes prefix length = x and stacking layer = y. Specifically, P=30 and N=3 achieve the best results on MVSA-S, while P=30 and N=5 are the optimal values for TumEmo.

To further analyze the influence of N and P, we also investigate the effect of changes in P or N on sentiment analysis when N or P is set to the optimal value.

Hyperparameter analysis of $\boldsymbol{N}$ with optimal $\boldsymbol{P}$. To further analyze the stacking layer N, we set P to optimal value (i.e., P is set to 30 on MVSA-S and TumEmo). The results illustrated in Fig 5 show that suitable stacking layer makes the model perform significantly. Besides, we also observe that small stacking layer (N=3) can obtain optimal performance on small dataset MVSA-S, while larger-scale dataset TumEmo requires larger stacking layer (N=5). This phenomenon suggests that it would be easier to obtain bidirectional sentiment interaction on small datasets. However, for the large-scale datasets, a deep stacked BIE is more convenient to obtain bidirectional interaction.

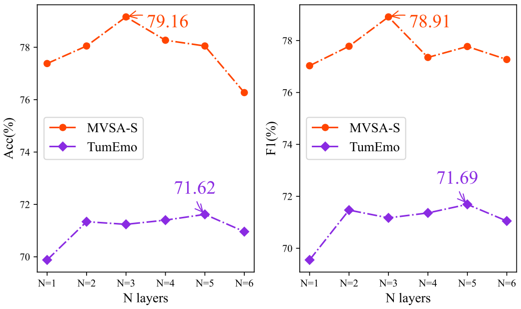


Fig. 5. The performance of N with optimal P.

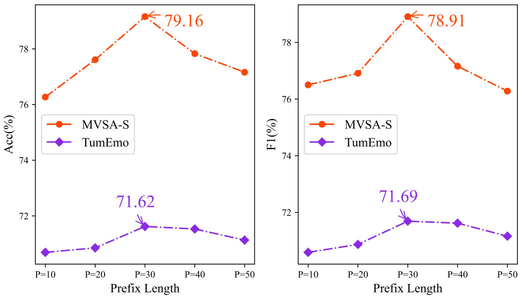


Fig. 6. The performance of P with optimal N.

Hyperparameter analysis of $\boldsymbol{P}$ with optimal $\boldsymbol{N}$. We also conduct experiments to analyze the prefix length P with optimal N (i.e., N is set to 3 and 5 on MVSA-S and TumEmo respectively). As shown in Fig 6, we observe that P=30 is the optimal value on both MVSA-S and TumEmo, which indicates that P=30 would convey more necessary information between FI and BI. Then, the conveyed information effectively improves the sentimental interaction learning of PIA.

Based on the above analysis, we set (P=30, N=3) and (P=30, N=5) on MVSA-S and TumEmo in our experiments.

G. Bidirectional Interaction Fusion Analysis

As shown in Fig 4, the scalar λ learned by AG indicates the importance of backward interaction and forward interaction during fusing them. Intuitively, the value of λ will influence the performance of BIT. Thus, to verify the effectiveness of our designed AG, we compare AG with different pre-defined λ ($\lambda \in [0.1, \ldots, 0.9]$) and give studies on some image-text pairs with learned λ.

The results of per-defined λ and BIT with AG are reported in Table V. We can find that different values of λ will get different results of sentiment analysis. Moreover, compared with the per-defined λ, BIT with adaptive λ learned by AG obtains the best performance. Some image-text pairs with adaptive λ learned by AG are shown in Fig 7, we observe that the values of learned λ are different for diverse image-text pairs. This illustrates that AG is able to adaptively learn the optimal λ for each image-text pair, which greatly improves the information integration during fusing f_o and b_o. This would

TABLE V
COMPARISON WITH PRE-DEFINED SETTINGS.

Model	MVSA-S		TumEmo	
	Acc	F1	Acc	F1
λ=0.1	76.94	76.01	70.80	70.45
λ=0.2	75.39	75.52	70.74	70.78
λ=0.3	77.16	76.85	70.96	70.83
λ=0.4	76.94	76.09	70.99	70.81
λ=0.5	76.94	76.02	71.05	70.99
λ=0.6	77.16	76.63	70.85	70.77
λ=0.7	77.83	77.43	71.23	71.04
λ=0.8	77.61	77.16	71.44	71.34
λ=0.9	78.05	77.25	71.32	71.09
BIT	**79.16**	**78.91**	**71.62**	**71.69**

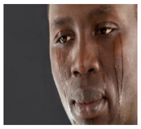
Image-text pair	λ	Image-text pair	λ
diego costa 'This is Who I am'. #passionate	0.87	 @Alexo670 I cant find my copy of skyrim #depressed	0.64
This is a very Manchester sky!	0.96	The colors of #borghetto and the gloomy sky!	0.72
the point when people argue so much, it becomes this # cute	0.59	<user> morning sunshine # grumpycat	0.93

Fig. 7. The image-text pairs with learned λ.

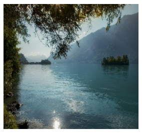
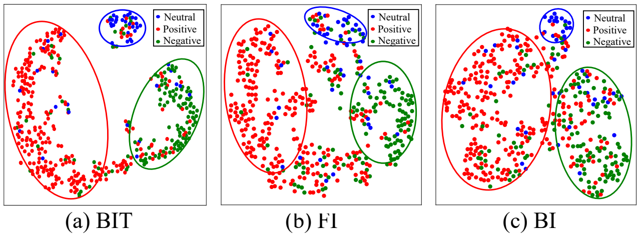
Image-text pair		GT	BIT	FI	BI
	the most beautiful lake in Switerland Brienzersee - with this gorgeous water color	Calm	Calm	Happy	Happy
	going to stores crippled like.	Neutral	Neutral	Negative	Positive

Fig. 8. Label prediction samples.



(a) BIT (b) FI (c) BI

Fig. 9. Cluster visualization on MVSA-S.

not be achieved by pre-defined λ. For example, despite the images in the second row of Fig 7 are highlighted by the region of "gloomy sky" and the corresponding word "sky" appears in the text, the learned λ are different. Clearly, if the λ is per-defined (e.g., λ=0.8), it will lead to getting suboptimal performance.

Based on the above observations, we believe that AG enhances the utilization of forward interaction and backward interaction information, which improves the performance of BIT.

H. Bidirectional Interaction Studies

Label prediction study. To further demonstrate the effectiveness of our model, we compare the sentiment/emotion label predicted by BIT, Forward Interaction branch (FI), and Backward Interaction branch (BI). The predicted results are shown in Fig 8, where GT denotes ground truth label. From the predicted results we can find that BIT achieves the best performance compared to FI and BI. This indicates that BIT models both forward and backward interaction and is able to convey interaction information between FI and BI allowing it to obtain more accurate predictions. For example, in the first row, BI and FI would focus more on the words including "beautiful lake" and "gorgeous water" and pay more attention to the water of the image, but they fail to capture the context of calm embedded in the whole image. However, since BIT models bidirectional interaction, it can capture the context of calm. Accordingly, BIT gets correct emotion prediction.

Cluster visualization study. In order to further verify that our proposed bidirectional interaction can learn more accurate sentiment interaction in image-text data, we conduct a cluster visualization study on MVSA-S. Specifically, we employ t-SNE to reduce the output of the models to 2-dimensional feature vector and visualize it. As shown in Fig 9, compared with FI and BI, the features extracted by BIT have better clusters in the feature space, which demonstrates that BIT can better distinguish the image-text sentiments embedded in image-text pairs. Thus, we believe that the bidirectional sentiment interaction learned by BIT significantly improves the performance of image-text sentiment analysis.

V. CONCLUSION

In this paper, we propose a novel bidirectional image-text interaction network BIT[1] for image-text sentiment analysis. Unlike previous work that used bidirectional information from unimodal data, BIT utilizes bidirectional information at the level of image-text interaction from image-text features. The experiments conducted on two public datasets demonstrate the effectiveness of our model.

While encouraging performance has been achieved, there are limitations of our model. For example, although BIE is a 4-input encoder, its inputs are extracted from image and text modalities, which limits the application of BIE in other multi-input tasks. In future, we would try to develop a network that can be applied on image-text and visual-text-audio sentiment analysis tasks.

[1] Code: https://github.com/hkxiaodong/BIT

References

[1] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A survey of sentiment analysis from social media data," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 450–464, 2020.

[2] W. Guo, Y. Zhang, X. Cai, L. Meng, J. Yang, and X. Yuan, "Ld-man: Layout-driven multimodal attention network for online news sentiment recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 1785–1798, 2021.

[3] S. Yang, Q. Li, W. Li, X. Li, and A.-A. Liu, "Dual-level representation enhancement on characteristic and context for image-text retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 8037–8050, 2022.

[4] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22, 2022, p. 2897–2905.

[5] P. Basu, S. Tiwari, J. Mohanty, and S. Karmakar, "Multimodal sentiment analysis of metoo tweets using focal loss (grand challenge)," in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 2020, pp. 461–465.

[6] A. Hu and S. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18, 2018, p. 350–358.

[7] N. Xu, "Analyzing multimodal public sentiment based on hierarchical semantic attentional network," in *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2017, pp. 152–154.

[8] N. Xu and W. Mao, "Multisentinet: A deep semantic network for multimodal sentiment analysis," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17, 2017, p. 2399–2402.

[9] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18, 2018, p. 929–932.

[10] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014–4026, 2021.

[11] X. Yang, S. Feng, Y. Zhang, and D. Wang, "Multimodal sentiment detection based on multi-channel graph neural networks," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online, Aug. 2021, pp. 328–339.

[12] X. Hu and M. Yamamura, "Two-stage attention-based fusion neural network for image-text sentiment classification," in *Proceedings of the 2022 4th International Conference on Image, Video and Signal Processing*, ser. IVSP '22, 2022, p. 1–7.

[13] G. Chen, J. Peng, W. Zhang, K. Huang, F. Cheng, H. Yuan, and Y. Huang, "A region group adaptive attention model for subtle expression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1613–1626, 2023.

[14] J. E. Steephen, S. C. Obbineni, S. Kummetha, and R. S. Bapi, "Hed-id: An affective adaptation model explaining the intensity-duration relationship of emotion," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 736–750, 2018.

[15] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3375–3385, 2023.

[16] J. Zhang, X. Liu, Z. Wang, and H. Yang, "Graph-based object semantic refinement for visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3036–3049, 2022.

[17] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[20] 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 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] Y. Zhang, Y. Zhang, W. Guo, X. Cai, and X. Yuan, "Learning disentangled representation for multimodal cross-domain sentiment analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7956–7966, 2023.

[22] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1966–1978, 2023.

[23] Z. Li, B. Xu, C. Zhu, and T. Zhao, "CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection," in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States, Jul. 2022, pp. 2282–2294.

[24] H. Zhu, Z. Zheng, M. Soleymani, and R. Nevatia, "Self-supervised learning for sentiment analysis via image-text matching," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1710–1714.

[25] P. Kumar, V. Khokher, Y. Gupta, and B. Raman, "Hybrid fusion based approach for multimodal emotion recognition with insufficient labeled data," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 314–318.

[26] L. Wang, W. Guo, X. Yao, Y. Zhang, and J. Yang, "Multimodal event-aware network for sentiment analysis in tourism," *IEEE MultiMedia*, vol. 28, no. 2, pp. 49–58, 2021.

[27] T. Jiang, J. Wang, Z. Liu, and Y. Ling, "Fusion-extraction network for multimodal sentiment analysis," in *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 785–797.

[28] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of The Web Conference 2020*, ser. WWW '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2514–2520.

[29] G. S. Cheema, S. Hakimov, E. Müller-Budack, and R. Ewerth, "A fair and comprehensive comparison of multimodal tweet sentiment analysis methods," in *Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding*, ser. MMPT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 37–45.

[30] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," *Know.-Based Syst.*, vol. 167, no. C, p. 26–37, mar 2019.

[31] A. Yadav and D. K. Vishwakarma, "A deep multi-level attentive network for multimodal sentiment analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 1, jan 2023.

[32] L. Xu, H. Zhang, L. Song, and Y. Lei, "Bi-mgan: Bidirectional t1-to-t2 mri images prediction using multi-generative multi-adversarial nets," *Biomedical Signal Processing and Control*, vol. 78, p. 103994, 2022.

[33] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2022.

[34] J. Devlin, M.-W. Chang, K. Lee, and 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, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[35] B. Kosko, "Bidirectional associative memories: Unsupervised hebbian learning to bidirectional backpropagation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 103–115, 2021.

[36] B. Wang, B. Yan, C. Liu, R. Hwangbo, G. Jeon, and X. Yang, "Lightweight bidirectional feedback network for image super-resolution," *Computers and Electrical Engineering*, vol. 102, p. 108254, 2022.

[37] D. Zhang, P. Huang, X. Ding, F. Li, W. Zhu, Y. Song, and G. Yang, "L2bec2: Local lightweight bidirectional encoding and channel attention cascade for video frame interpolation," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2, pp. 1–19, 2023.

[38] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[40] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multiview social data," in *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*. Springer, 2016, pp. 15–27.

[41] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.