

# Hierarchical Interactive Multimodal Transformer for Aspect-Based Multimodal Sentiment Analysis

Jianfei Yu<sup>✉</sup>, Kai Chen, and Rui Xia

**Abstract**—Aspect-based multimodal sentiment analysis (ABMSA) aims to determine the sentiment polarities of each aspect or entity mentioned in a multimodal post or review. Previous studies to ABMSA can be summarized into two subtasks: aspect-term based multimodal sentiment classification (ATMSC) and aspect-category based multimodal sentiment classification (ACMSC). However, these existing studies have three shortcomings: (1) ignoring the object-level semantics in images; (2) primarily focusing on aspect-text and aspect-image interactions; (3) failing to consider the semantic gap between text and image representations. To tackle these issues, we propose a general Hierarchical Interactive Multimodal Transformer (HIMT) model for ABMSA. Specifically, we extract salient features with semantic concepts from images via an object detection method, and then propose a hierarchical interaction module to first model the aspect-text and aspect-image interactions, followed by capturing the text-image interactions. Moreover, an auxiliary reconstruction module is devised to largely eliminate the semantic gap between text and image representations. Experimental results show that our HIMT model significantly outperforms state-of-the-art methods on two benchmarks for ATMSC and one benchmark for ACMSC.

**Index Terms**—Fine-grained opinion mining, aspect-based sentiment analysis, multimodal sentiment analysis

## 1 INTRODUCTION

RECENT years have witnessed the rapid growth of user-generated contents on various online platforms. As these user-generated contents primarily reflect users' personal opinions, it becomes increasingly important to analyze them to identify public opinions towards target aspects or entities. Aspect-based sentiment analysis (ABSA) is the task of detecting sentiment polarities towards given aspects or entities in the input text. As an important task in sentiment analysis, ABSA has attracted much attention from both academia and industry during the last decade [1].

In the literature, many approaches have been proposed for ABSA, including traditional feature-based models [2], [3], [4], [5] and deep learning-based models [6], [7], [8], [9], [10], [11]. With the recent trend of fine-tuning pre-trained models in NLP tasks, a few studies attempted to apply the pre-trained BERT model [12] to ABSA and obtained the state-of-the-art performance on several benchmark datasets [13], [14], [15]. In spite of the remarkable advancement, most of these approaches heavily rely on the textual contents, failing to consider other associated modalities (e.g., facial expressions in images). As many online platforms have become increasingly multimodal, the information from other modalities is also

important for predicting the sentiment polarities over target aspects. Inspired by this, a number of recent studies proposed to leverage the useful information from images to improve the performance of the ABSA task [16], [17], [18].

Following these recent studies, in this paper, we focus on aspect-based multimodal sentiment analysis (ABMSA), which aims to identify the sentiment orientation over each aspect or entity given a multimodal post or review. Current works on ABMSA are primarily centered around its two subtasks:

- Aspect-Term based Multimodal Sentiment Classification (ATMSC), where the goal is to predict the sentiment over each target entity, which could be a single word or a phrase mentioned in the input text. For example, in Fig. 1a, the task of ATMSC requires inferring the sentiment polarity over the two entities "SamHunt" and "Stagecoach" mentioned by the multimodal tweet.
- Aspect-Category based Multimodal Sentiment Classification (ACMSC), where the goal is to detect the sentiment with regard to each pre-defined aspect category. For example, in Fig. 1b, the task of ACMSC requires inferring the sentiment ratings over the pre-defined aspect categories based on the multimodal review.

As the two subtasks of ABMSA, their key difference is the aspect definition, i.e., whether the input aspect is a pre-defined category or a term mentioned in the input text, but the overall setup of the two tasks are essentially the same. Therefore, our objective in this work is to propose a general multimodal architecture to encompass both tasks of ATMSC and ACMSC.

Although previous studies have demonstrated the effectiveness of incorporating visual information into ABSA [16], [17], [18], they still suffer from the following limitations: (1) These methods only utilize the hidden representations of input

- The authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: {jfyu, kchen, rxia}@njjust.edu.cn.

Manuscript received 19 June 2021; revised 19 Apr. 2022; accepted 24 Apr. 2022. Date of publication 28 Apr. 2022; date of current version 13 Sept. 2023.

This work was supported in part by the Natural Science Foundation of China under Grants 62076133 and 62006117, and in part by the Natural Science Foundation of Jiangsu Province for Young Scholars under Grant BK20200463 and in part by Distinguished Young Scholars under Grant BK20200018. (Corresponding author: Rui Xia.)

Recommended for acceptance by C. Busso.

Digital Object Identifier no. 10.1109/TAFFC.2022.3171091

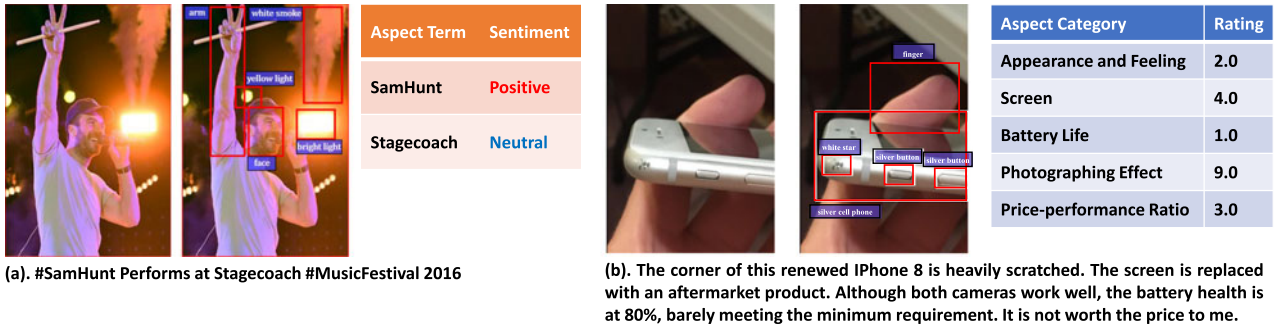


Fig. 1. Examples for Aspect-Based Multimodal Sentiment Analysis (ABMSA). Fig. 1a shows an example for aspect-term based multimodal sentiment classification (ATMSC), where the goal is to detect the positive and neutral sentiment towards the two entities “SamHunt” and “Stagecoach,” respectively; while Fig. 1b shows an example for aspect-category based multimodal sentiment classification (ACMSC), where the goal is to detect the sentiment rating for each pre-defined aspect category.

images, but ignore the object-level semantic information in them. These semantic information can potentially guide the model to focus on image regions related to the given aspect, and alleviate the noise from irrelevant regions (e.g., the middle images in Figs. 1a and 1b). (2) Existing methods primarily focus on modeling aspect-image and aspect-text interactions, but fail to pay enough attention to the information flow between text and images. Actually, the image can often guide the model to focus on aspect-related sentiment words in text. For example, in Fig. 1a, with *SamHunt* as the input aspect, the image object with smiling face may help highlight the word *Festival* in text; in Fig. 1b, with *appearance and feeling* as the input aspect, the image object containing the *white star* may help highlight *heavily scratched* in text. (3) In previous studies, although the text representation and the image representation are usually obtained from different pre-trained models, they ignore the semantic gap between the text and the image representations, which may increase the risk of misalignment in their inter-modal interactions.

To address the aforementioned limitations, we propose a general multimodal architecture named Hierarchical Interactive Multimodal Transformer (HIMT) for ABMSA. As illustrated in Fig. 2, our main contributions can be summarized as follows:

- To strengthen the semantic meanings of image representations in Unimodal Feature Extraction Module,

we propose to detect a set of salient objects in each image based on a pre-trained Faster R-CNN model, and represent each object by concatenating its hidden representation and associated semantic concepts, followed by an Aspect-Guided Attention layer to learn the relevance of each semantic concept with the guidance of given aspects.

- To model the pairwise interactions between the given aspects, the text, and the images, we propose a Hierarchical Interaction Module, which contains an Aspect-Aware Transformer layer to obtain aspect-aware text representations and aspect-aware image representations at the lower level, and a carefully-designed Multimodal Fusion Transformer layer to capture the inter-modal interactions between the text and the image representations at the higher level.
- Finally, to bridge the semantic gap between the text representation and the image representation, we further propose an auxiliary reconstruction module based on the idea of auto-encoder [19] to reconstruct the input textual context from these two representations respectively, and successfully gain extra improvements.

Experimental results demonstrate that our HIMT model consistently performs better than a number of unimodal and multimodal approaches including a recently pre-trained multimodal model, and refresh the state-of-the-art results on two

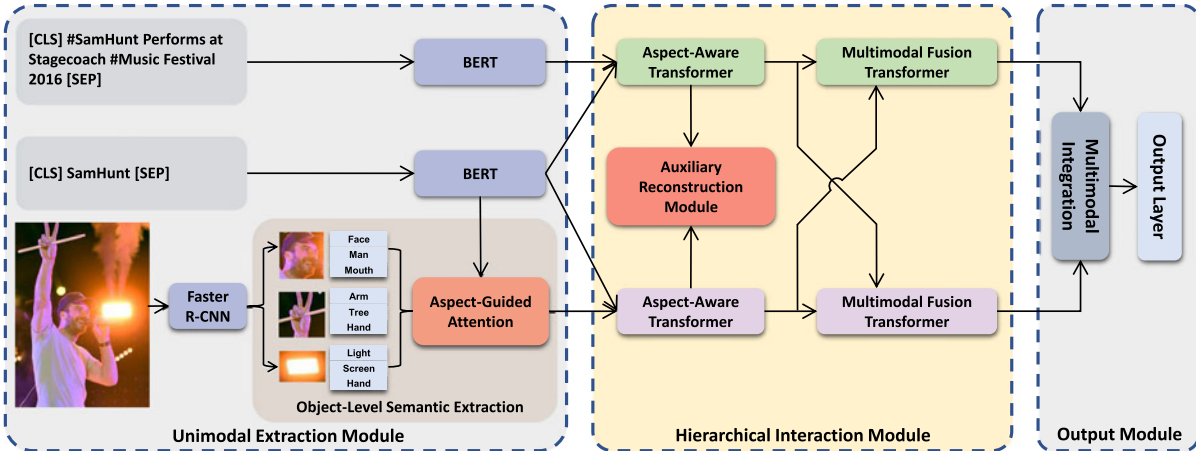


Fig. 2. Overview of Our Hierarchical Interactive Multimodal Transformer (HIMT). HIMT consists of four modules: Unimodal Feature Extraction Module, Hierarchical Interaction Module, Auxiliary Reconstruction Module, and Output Module.

benchmark datasets for ATMSC and one benchmark dataset for ACMSC.<sup>1</sup>

## 2 RELATED WORK

### 2.1 Aspect-Based Sentiment Analysis (ABSA)

As an important fine-grained sentiment analysis task, ABSA has been well studied in the past decade [1], [20]. Traditional methods focus on manually designing many aspect-specific features, and training a statistical learning classifier based on these features for sentiment classification [3], [4], [5], [21]. With recent advancements in deep learning, a large number of studies adopt various neural network models to encode the input aspect and the related context, including Recursive Neural Network [22], Convolutional Neural Network [23], Recurrent Neural Network [24], [25], [26]. Besides, different kinds of attention mechanisms have been introduced to model the interactions between the aspect and the context [6], [7], [8], [15], [27], [28], [29]. Moreover, Xu *et al.* [13] and Sun *et al.* [14] recently showed the success of applying the pre-trained BERT model to this task, which have achieved the state-of-the-art results on multiple benchmark datasets.

Despite obtaining remarkable performance, the aforementioned approaches only focus on the textual modality, but failing to consider the associated information from other modalities. To address this, Xu *et al.* [16] explored the task of aspect category-based multimodal sentiment classification (ACMSC), where they first created a multimodal review dataset, and then proposed a Multi-Interactive Memory Network. Yu *et al.* [18] and Yu and Jiang [17] studied the task of aspect term-based multimodal sentiment classification (ATMSC), where they manually constructed two Twitter datasets, followed by proposing an entity-sensitive attention and fusion network and a target-oriented multimodal BERT architecture, respectively. In this paper, we aim to extend this line of research by proposing a more effective multimodal method for both ACMSC and ATMSC tasks.

### 2.2 Multimodal Sentiment Analysis (MSA)

As an increasingly popular area of affective computing research, MSA aims to integrate language and other nonverbal modalities (e.g., vision and acoustic modality) to detect the user sentiment [30]. Most existing studies in MSA can be summarized into two groups: conversational MSA and social media MSA.

In conversational MSA, many existing multimodal approaches focus on (1) adapting different shallow and deep models to the task, including Support Vector Machine, Convolutional Neural Network, Long Short-Term Memory Network, Gated Recurrent Unit, and Transformer, which have been shown to achieve satisfactory performance on many MSA tasks such as emotion classification [31], [32], sentiment classification [30], [33], [34], [35], and sarcasm detection [36], [37]; and (2) building the interactions between different modalities based on different fusion paradigms [38], including early and late fusion [33], tensor fusion [30], [39], multimodal gated units [40], [41], and temporal attention mechanism [42], [43]. However, these

methods are mainly designed for sentiment analysis in conversations, and cannot be directly applied to aspect-level multimodal sentiment analysis. Moreover, most of them are based on traditional neural networks, while our work aims to construct a multimodal method based on the recent pre-trained BERT model.

In social media MSA, one line of work focuses on proposing various effective models to perform visual sentiment analysis of social images [44], [45], [46], [47], [48]. Another line of work aims to combine the textual and image information to perform tweet-level sentiment analysis of multimodal social posts [49], [50]. Different from these studies that focus on coarse-grained MSA (i.e., detecting the overall sentiment of each social post), our objective in this paper is to perform fine-grained MSA (i.e., detecting the sentiment over all the aspects mentioned in each social post) by integrating textual and visual inputs.

## 3 METHODOLOGY

In this section, we first formulate our task, and introduce the overall architecture of our Hierarchical Interactive Multimodal Transformer (HIMT). We then delve into the details of each module in HIMT.

**Task Formulation.** Given a multimodal post or review as input, the goal of aspect term-based multimodal sentiment classification (ATMSC) is to infer the sentiment over the given target entities, and the goal of aspect category-based multimodal sentiment classification (ACMSC) is to infer the sentiment with respect to the pre-defined aspect category. Note that in this paper, *sentiment* refers to users' opinions or attitudes towards named entities (e.g., Person, Location, and Organization) in social posts or product aspects (e.g., screen, price, and appearance) in reviews.

Formally, for each sample, we are given a piece of text with  $m$  words  $S = (w_1, w_2, \dots, w_m)$ , a set of  $l$  associated images  $I = (V_1, V_2, \dots, V_l)$ , and  $r$  aspects  $(T_1, T_2, \dots, T_r)$ . With a  $(S, I)$  pair and one of the aspects  $T_j$  as input, our goal is to predict the sentiment orientation  $y \in \mathcal{Y}$  over the given aspect  $T_j$ . In this paper, for ATMSC,  $\mathcal{Y}$  denotes the label set containing three classes, namely *positive*, *negative*, and *neutral*; while for ACMSC,  $\mathcal{Y}$  denotes the sentiment rating set ranging from 1 to 10.

### 3.1 Overview of Our Proposed Approach

Fig. 2 illustrates the overall architecture of HIMT, which contains the following four modules: (1) The Unimodal Feature Extraction Module focuses on extracting high-level feature representations from the input aspect, the input text, and the input image, respectively. (2) On basis of this, a Hierarchical Interaction Module is designed to first model the aspect-text and aspect-image interactions to obtain aspect-aware text representations (ATR) and aspect-aware image representations (AIR), followed by capturing the inter-modal dynamics between ATR and AIR to obtain multimodal-fused text/image representations. (3) To bridge the semantic gap between ATR and AIR, we further propose an Auxiliary Reconstruction Module to minimize their representation divergence by reconstructing the input text representation from ATR and AIR, respectively. (4) Finally, the

1. We will make our source codes publicly available via Github.

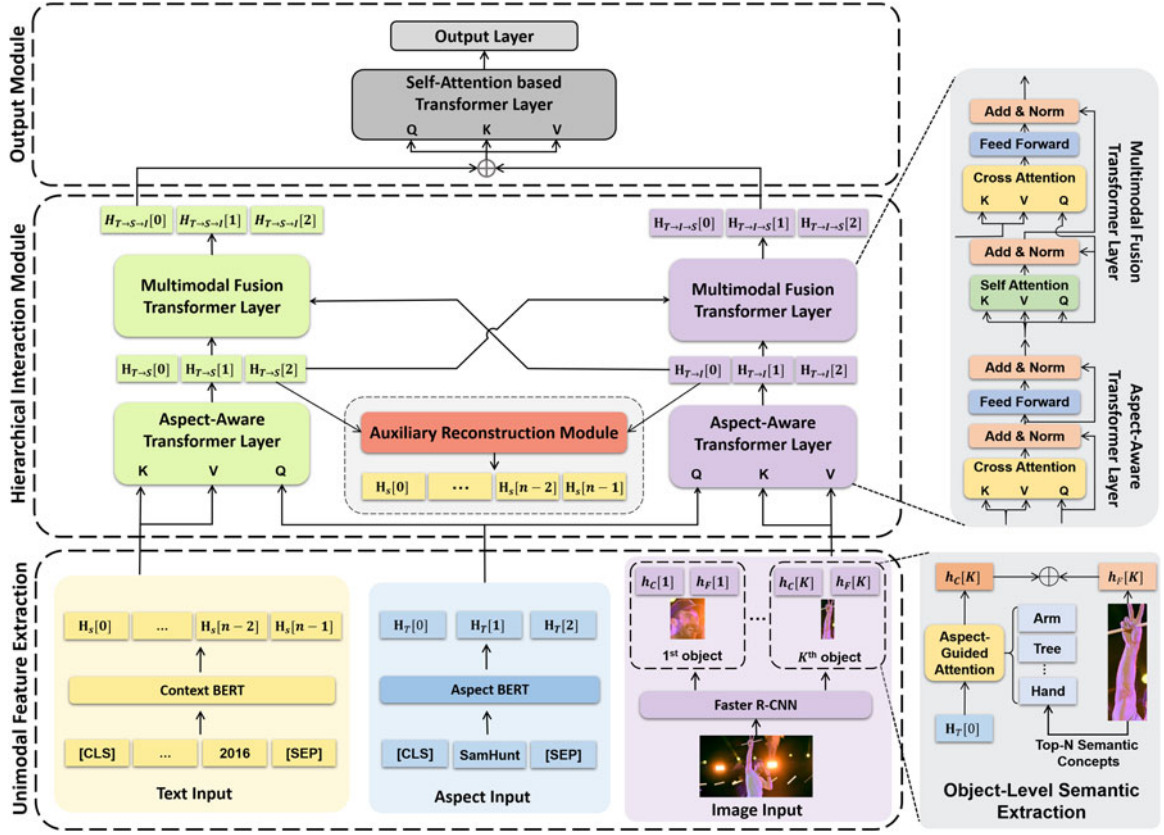


Fig. 3. The architecture of our Hierarchical Interactive Multimodal Transformer (HIMT) model. The two sub-figures in the right illustrate the model details in Object-Level Semantic Extraction and Hierarchical Interaction Module, respectively.

integrate the multimodal-fused text and image representations, followed by a softmax layer for sentiment predictions.

### 3.2 Unimodal Feature Extraction Module

We first adopt two pre-trained models to extract high-level feature representations from the aspect, the text, and the image respectively.

#### 3.2.1 Aspect and Text Representation With BERT

Following many previous works in Textual Aspect-Based Sentiment Analysis [27], [51], we feed the aspect and the input text into two textual encoders in order to better differentiate the input sentence with multiple aspect terms or aspect categories. Specifically, for the input aspect  $T$ , we follow the practice in BERT [12] by appending a special token [CLS] to the beginning and another special token [SEP] to the end, and employ a pre-trained BERT encoder to get its hidden representation, denoted by  $\mathbf{H}_T = \text{BERT}(T)$ , where  $\mathbf{H}_T \in \mathbb{R}^{t \times d}$  is the generated aspect representation,  $d$  is the hidden dimension, and  $t$  is the length of the aspect.

For the input text  $S$ , we insert two special tokens (i.e., [CLS] and [SEP]) to the beginning and the end, respectively, and use another pre-trained BERT model to obtain the hidden representation:  $\mathbf{H}_S = \text{BERT}(S)$ , where  $\mathbf{H}_S \in \mathbb{R}^{n \times d}$  is the generated text representations and  $n$  is the length of  $S$ . Note that for the task of ATMSC, since the given aspect term (i.e., target entity) is a word or a phrase within  $S$ , we replace the aspect with a special token  $\$T\$$

to indicate its position in  $S$ , and use the modified text as  $S$ . Take Fig. 1a as an example. Its text and aspect inputs are shown in the bottom of Fig. 3.

#### 3.2.2 Image Representation With Object-Level Semantic Extraction

Although some recent multimodal studies have employed different pre-trained image encoders to obtain object-level visual features [52], [53], these studies only consider the hidden representation of each object, but ignore the semantic information contained by these objects, which may help strengthen the semantic meanings of image representations. Therefore, as shown in the right bottom of Fig. 3, we propose to represent each object by combining its hidden representation and corresponding semantic information.

**Faster R-CNN.** As one of the state-of-the-art object detection methods, Faster R-CNN [54] has been demonstrated to achieve satisfactory performance in identifying objects in images. It mainly consists of two key steps: the first step is to detect object proposals with Region Proposal Network (RPN), and the second step is to predict the object category (e.g., man) as well as its associated attribute (e.g., smiling) of each object proposal. Specifically, given an image  $V$  in the input image set  $I$ , we first adopt the Faster R-CNN model pre-trained on Visual Genome dataset [55] to extract all the object proposals, and then only keep  $K$  objects with the highest confidence to avoid the noise brought by less useful low-scoring



objects. Formally, we denote the detected object and its corresponding semantic meaning as follows:

$$(h_{F_i}, C_i) = \text{Faster R-CNN}(V), \quad i = 1, 2, \dots, K, \quad (1)$$

where  $h_{F_i} \in \mathbb{R}^{2048}$  is the mean-pooled convolutional feature of the  $i$ th object in  $V$  and  $C_i$  is the top one of the pre-defined 1,600 object categories predicted by Faster R-CNN.<sup>2</sup>

**Top- $N$  Semantic Concepts.** Since the predicted object category  $C_i$  will lead to the error propagation issue if it is wrongly predicted, we propose to keep Top- $N$  predicted categories for each object, denoted by  $C_i = \{c_1, c_2, \dots, c_N\}$ . Since each object category  $c_j$  is a description with one or multiple tokens (e.g., *smiling face*), we employ the average pooling operation over word embeddings of all the tokens to obtain the representation of  $c_j$ , denoted by  $h_{c_j} = \frac{1}{m} \sum_{k=1}^m w_{jk}$ , where  $m$  denotes the number of tokens, and  $w_{jk} \in \mathbb{R}^{300}$  refers to the word embedding of the  $k$ th token retrieved from GloVe [56].

**Aspect-Guided Attention Layer.** Although incorporating Top- $N$  categories for each object can alleviate the error propagation issue, it will also bring some noise. For example, in the right part of Fig. 3, the Top- $N$  categories for the hand posture object contains a wrong category *tree*, which is irrelevant to the input aspect *SamHunt* and may lead our model to ignore the hand posture object. To alleviate the noise from unrelated object categories, we use an aspect-guided attention mechanism to dynamically learn the relevance between the input aspect and each predicted object category.

Specifically, we treat the hidden representation of the [CLS] token in the aspect as the aspect representation, i.e.,  $h_T = \mathbf{H}_T^{[\text{CLS}]}$ , and use it as a query vector to generate the attention weights for all the Top- $N$  categories

$$e_j = h_T \mathbf{W}_a^T h_{c_j} \quad (2)$$

$$\alpha_j = \frac{\exp(e_j)}{\sum_{k=1}^N \exp(e_k)} \quad (3)$$

where  $\mathbf{W}_a \in \mathbb{R}^{300 \times d}$  is the weight parameter. Based on the attention weights, we can obtain the weighted representation of the Top- $N$  categories for the  $i$ th object  $C_i$  as follows:

$$h_{C_i} = \sum_{j=1}^N \alpha_j h_{c_j}. \quad (4)$$

**Final Image Representation.** Next, we concatenate the semantic category representation  $h_{C_i}$  and the hidden representation  $h_{F_i}$  in Eqn. (1) to obtain the final representation of the  $i$ th object

$$h_{V_i} = [h_{F_i}; h_{C_i}], \quad (5)$$

where  $h_{V_i} \in \mathbb{R}^{2048+300}$ . Finally, we concatenate the representations of the detected  $K$  objects as the final image representation, and then project it into the same dimension of the text representation

$$\mathbf{H}_V = \mathbf{W}^T [h_{V_1}; h_{V_2}; \dots; h_{V_K}], \quad (6)$$

where  $\mathbf{W} \in \mathbb{R}^{(2048+300) \times d}$  is a weight parameter.

It is worth noting that if the input image set  $I$  has only one image  $V$ , we treat  $\mathbf{H}_V$  as its final image representation  $\mathbf{H}_I$ ; otherwise, we concatenate the representations of all the  $l$  images in  $I$  as its final representation

$$\mathbf{H}_I = [\mathbf{H}_{V_1}; \mathbf{H}_{V_2}; \dots; \mathbf{H}_{V_l}]. \quad (7)$$

### 3.3 Hierarchical Interaction Module

After obtaining the text and image representations, we propose a Hierarchical Interaction Module, which contains two levels of interactions: (1) the lower-level interactions between aspects and texts (or images); (2) the higher-level interactions between texts and images. As illustrated in the intermediate part of Fig. 3, we will present the details of these two interactions in the following subsections.

#### 3.3.1 Aspect-Aware Transformer Layer

To model the interaction between aspects and texts as well as the interaction between aspects and images, we introduce an Aspect-Aware Transformer (AAT) layer based on the Cross-Modal Transformer layer proposed by Tsai *et al.* [43]. Specifically, as illustrated in the left part of Fig. 3, we first adopt a multi-head cross attention (MCATT) mechanism, which treats the aspect representation  $\mathbf{H}_T$  as queries, and the text representation  $\mathbf{H}_S$  as keys and values, followed by two layer normalization [57] (LN) and a feed-forward network [58] (FFN) as follows:

$$\mathbf{Z} = \text{LN}(\mathbf{H}_T + \text{MCATT}(\mathbf{H}_T, \mathbf{H}_S)) \quad (8)$$

$$\mathbf{H}_{T \rightarrow S} = \text{LN}(\text{FFN}(\mathbf{Z}) + \mathbf{Z}), \quad (9)$$

where  $\mathbf{H}_{T \rightarrow S} \in \mathbb{R}^{t \times d}$  is the generated aspect-aware text representation of the AAT layer, i.e.,  $\text{AAT}(\mathbf{H}_T, \mathbf{H}_S)$ . To obtain a deeper interaction between aspects and texts, we further propose to stack  $l$  AAT layers as follows:

$$\mathbf{H}_{T \rightarrow S}^{[l]} = \text{AAT}(\mathbf{H}_{T \rightarrow S}^{[l-1]}, \mathbf{H}_S), \quad (10)$$

where  $\mathbf{H}_{T \rightarrow S}^{[l]} \in \mathbb{R}^{t \times d}$  is the final aspect-aware text representation at the  $l$ th layer.

Similar to  $\mathbf{H}_{T \rightarrow S}^{[l]}$ , as shown in the right part of Fig. 3, we utilize the same structure to obtain the aspect-aware image representation  $\mathbf{H}_{I \rightarrow I}^{[l]} \in \mathbb{R}^{t \times d}$

$$\mathbf{H}_{I \rightarrow I}^{[l]} = \text{AAT}(\mathbf{H}_{I \rightarrow I}^{[l-1]}, \mathbf{H}_I). \quad (11)$$

#### 3.3.2 Multimodal Fusion Transformer Layer

To further capture the inter-modal interaction between the aspect-aware text and image representations, we propose a Multimodal Fusion Transformer (MFT) Layer on top of the AAT layers. Note that an intuitive solution to capture the inter-modal interaction is to employ the Cross-Modal Transformer (CMT) layer [43]. However, simply employing the CMT layer will inevitably overemphasize the higher-level

2. In Faster R-CNN, the predicted object category may contain the object attribute, when its prediction confidence is higher than a pre-defined threshold.

text-image interaction and pay less attention to the important lower-level aspect-aware text/image representations derived from the AAT layers. Therefore, the MFT layer is proposed to not only model the higher-level inter-modal interaction via a CMT layer but also strengthen the lower-level aspect-aware text/image representations via a self-attention layer.

Specifically, as shown in the top left part of Fig. 3, we first feed the  $\mathbf{H}_{T \rightarrow S}^{[l]}$  to a multi-head self-attention layer with layer normalization to strengthen the aspect-aware text representation

$$\tilde{\mathbf{H}}_{T \rightarrow S} = \text{LN}\left(\mathbf{H}_{T \rightarrow S}^{[l]} + \text{MSATT}\left(\mathbf{H}_{T \rightarrow S}^{[l]}\right)\right), \quad (12)$$

where MSATT refers to the multi-head self-attention layer, and  $\tilde{\mathbf{H}}_{T \rightarrow S}$  is the output.

To capture the inter-modal interactions, we employ a CMT layer, where  $\mathbf{H}_{T \rightarrow S}^{[l]}$  is treated as the query and  $\mathbf{H}_{T \rightarrow I}^{[l]}$  is treated as keys and values

$$\tilde{\mathbf{H}}_{T \rightarrow S \rightarrow I} = \text{MCATT}\left(\mathbf{H}_{T \rightarrow S}^{[l]}, \mathbf{H}_{T \rightarrow I}^{[l]}\right). \quad (13)$$

Next, to integrate  $\tilde{\mathbf{H}}_{T \rightarrow S}$  and  $\tilde{\mathbf{H}}_{T \rightarrow S \rightarrow I}$ , we propose to feed  $\tilde{\mathbf{H}}_{T \rightarrow S \rightarrow I}$  to a FFN layer followed by a LN layer with residual connections from  $\tilde{\mathbf{H}}_{T \rightarrow S}$

$$\mathbf{H}_{T \rightarrow S \rightarrow I} = \text{LN}(\text{FFN}(\tilde{\mathbf{H}}_{T \rightarrow S \rightarrow I}) + \tilde{\mathbf{H}}_{T \rightarrow S}), \quad (14)$$

where  $\mathbf{H}_{T \rightarrow S \rightarrow I}$  is the output of MFT layer. Next, we further stack  $k$  such MFT layers to obtain the final multimodal-fused image representation, denoted by  $\mathbf{H}_{T \rightarrow S \rightarrow I}^{[k]} \in \mathbb{R}^{t \times d}$ .

Similar to  $\mathbf{H}_{T \rightarrow S \rightarrow I}^{[k]}$ , as shown in the top right part of Fig. 3, we also utilize the same structure by treating  $\mathbf{H}_{T \rightarrow I}^{[l]}$  as queries and  $\mathbf{H}_{T \rightarrow S}^{[l]}$  as keys and values, which helps obtain the multimodal-fused text representation, denoted by  $\mathbf{H}_{T \rightarrow I \rightarrow S}^{[k]} \in \mathbb{R}^{t \times d}$ .

### 3.4 Auxiliary Reconstruction Module

As mentioned before, since the text and image representations are respectively obtained from two pre-trained unimodal models (i.e., BERT and Faster R-CNN), they lie in different representation spaces, which leads to the cross-modal semantic gap. To largely eliminate the semantic gap between text and image representations, we devise an auxiliary reconstruction module in this subsection.

Specifically, to project the two representations into a shared space, we construct our auxiliary reconstruction model on basis of the idea from auto-encoder [19], which first projects the text and image inputs into intermediate representations with a shared encoder (i.e., aspect-aware transformer layer), and then reconstructs the inputs from the intermediate representations with a shared decoder (i.e., Auxiliary Reconstruction Module). Formally, after obtaining the aspect-aware text representation (i.e.,  $\mathbf{H}_{T \rightarrow S}^{[l]}$ ) and the aspect-aware image representation (i.e.,  $\mathbf{H}_{T \rightarrow I}^{[l]}$ ) via the shared encoders defined in Eqns. (10) and (11), we feed them to two separate fully-connected layers as follows:

$$\begin{aligned} \mathbf{H}_{S \rightarrow TS} &= \mathbf{W}_r \mathbf{H}_{T \rightarrow S}^{[l]} \\ \mathbf{H}_{S \rightarrow TI} &= \mathbf{W}_r \mathbf{H}_{T \rightarrow I}^{[l]}, \end{aligned} \quad (15)$$

where  $\mathbf{W}_r \in \mathbb{R}^{n \times t}$  is the weight parameter shared for the two fully-connected layers.

Next, we propose to utilize the Mean Squared Error (MSE) to minimize the divergence between each reconstructed text representation in Eqn. (15) and the original text representation obtained from BERT (i.e.,  $\mathbf{H}_S$ ) as the auxiliary objective function below

$$\mathcal{L}^r = \text{MSE}(\mathbf{H}_{S \rightarrow TS} - \mathbf{H}_S) + \text{MSE}(\mathbf{H}_{S \rightarrow TI} - \mathbf{H}_S). \quad (16)$$

### 3.5 Output Module

Finally, we treat the hidden representation of the first token (i.e., [CLS]) in  $\mathbf{H}_{T \rightarrow S \rightarrow I}^{[k]}$  as the final image representation, and concatenate it with the multimodal-fused text representation  $\mathbf{H}_{T \rightarrow I \rightarrow S}^{[k]}$ , followed by feeding it to a Transformer layer:

$$\mathbf{H} = \text{Transformer}\left(\mathbf{H}_{T \rightarrow S \rightarrow I}^{[k], [\text{CLS}]}, \mathbf{H}_{T \rightarrow I \rightarrow S}^{[k]}\right), \quad (17)$$

where  $\mathbf{H} \in \mathbb{R}^{(m+1) \times d}$  is the final multimodal representation generated from the Transformer layer.

*Output Layer.* We then feed the final hidden representation of the first token  $\mathbf{H}^{[0]}$  to a softmax layer for aspect-based sentiment classification as follows:

$$p(\mathbf{y}|\mathbf{H}) = \text{softmax}\left(\mathbf{W}_{\text{MM}}^\top \mathbf{H}^{[0]}\right), \quad (18)$$

where  $\mathbf{W}_{\text{MM}} \in \mathbb{R}^{d \times 3}$  is the weight matrix for ATMSC, and  $\mathbf{W}_{\text{MM}} \in \mathbb{R}^{d \times 10}$  is the weight matrix for ACMSC.

*Final Objective Function.* To optimize all the parameters in our HIMT model, the objective is to minimize the combined loss containing two parts. Given a set of training samples  $\mathcal{D}$ , the first part is the cross-entropy loss for the final prediction, and the second part is the weighted MSE loss of the reconstruction module

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \left( \sum_{j=1}^{|\mathcal{Y}|} t_{ij} \log p(y_{ij}|\mathbf{H}_i) + \lambda \mathcal{L}_i^r \right), \quad (19)$$

where  $|\mathcal{Y}|$  is 3 for ATMSC and 10 for ACMSC,  $t_{ij}$  is the ground-truth label, and  $\lambda$  is the tradeoff hyper-parameter for controlling the contribution of the reconstruction module.

## 4 EXPERIMENT

In this section, we conduct both extrinsic and intrinsic evaluations to show the effectiveness of our Hierarchical Interactive Multimodal Transformer (HIMT) model.

### 4.1 Experiment Setting

*Datasets.* As shown in Tables 1 and 2, we adopt two benchmark datasets TWITTER-2015 and TWITTER-2017 from Yu and Jiang [17] for ATMSC and one benchmark datasets ZOL from Xu *et al.* [16] for ACMSC. The two TWITTER datasets contain multimodal tweets posted during 2014-2015 and 2016-2017 on Twitter. The ZOL dataset is a Chinese

TABLE 1  
The Statistic of Two TWITTER Datasets

Label	TWITTER-2015			TWITTER-2017		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	517	573
Negative	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234

dataset collected from the reviews of some popular mobile phones in ZOL, which is the leading IT information web portal in China.

For the TWITTER datasets, all the aspect terms (i.e., target entities) were annotated by Zhang *et al.* [59] and Lu *et al.* [60], respectively. Given the manually labeled aspect terms, Yu and Jiang [17] further annotated the sentiment polarities over them. The basic statistics of the TWITTER datasets are shown in Table 1. Following Yu and Jiang [17], we use Accuracy (ACC) and Macro-F1 (F1) as our evaluation metrics.

For the ZOL dataset, each review consists of a text review and one or multiple associated images. Besides, it pre-defines a set of aspect categories, including price-performance ratio, performance configuration, battery life, appearance and feeling, photographing effect, and screen. Each aspect category has an associated sentiment rating from 1 to 10. Following Xu *et al.* [16], we use Accuracy (ACC) and Weighted-F1 (F1) as our evaluation metrics.

**Hyper-Parameters.** For the textual feature extraction, we adopt the pre-trained uncased English BERT<sub>base</sub> model [12] and Chinese BERT<sub>wwm</sub> [61] to obtain the aspect and text representations for the two TWITTER datasets and the ZOL dataset, respectively. For image feature extraction, we employ the Faster R-CNN [54] with ResNet-101 [62] backbone released by Anderson *et al.* [55]. The Faster R-CNN model is pre-trained on Visual Genome dataset [63], which contains 1,600 object categories and 400 attributes. The object proposals produced by Faster R-CNN are sorted by the object detection probabilities, and we select the top-36 objects as our image features. For each detected object, we employ the Top-3 categories as its semantic concepts, i.e.,  $N = 3$ . Besides, we adopt Glove [56] and SGNS [64] to obtain the English and Chinese word embeddings to encode the semantic concepts. Moreover, all the other hyper-parameters are set after tuning them on the development set. The values of hyper-parameters including learning rate, warm up proportion, trade-off parameter  $\lambda$ , number of AAT layers  $l$ , number of MFT layers  $k$ , maximum context and aspect lengths, training batch size, number of attention

TABLE 2  
The Statistic of the ZOL Dataset

Label (ratings)	ZOL		
	Train	Dev	Test
1, 2, 3, 4,	1152	148	139
5, 6, 7,	4670	588	581
8, 9, 10	16921	2107	2123
Total	22743	2843	2843

TABLE 3  
The Hyper-Parameter Setting of Our HIMT Model

Hyper-parameters	TWITTER-2015	TWITTER-2017	ZOL
learning rate	1e-5	2e-5	3e-5
warm up proportion	0.1	0.1	0.1
trade-off parameter $\lambda$	1e-2	1e-2	1e-2
number of AAT layers $l$	2	4	2
number of MFT layers $k$	1	1	1
maximum context length	64	64	128
maximum aspect length	16	16	16
training batch size	32	32	32
number of attention heads	12	12	12
hidden dimension $d$	768	768	768
number of training epochs	8	8	8

heads, hidden dimension  $d$ , and number of training epochs are shown in Table 3. All the models are implemented with PyTorch, and run on an NVIDIA TITAN RTX GPU.

## 4.2 Extrinsic Comparisons

In this subsection, we perform extrinsic evaluation by comparing our HIMT model with a number of existing competitive methods. Specifically, we consider the following unimodal and multimodal methods:<sup>3</sup>

- ResNet-Aspect: a baseline method that extracts the image and the aspect features with ResNet and BERT, respectively.
- Faster R-CNN-Aspect: another baseline that is similar to ResNet-Aspect but the image features are extracted by Faster R-CNN.
- AE-LSTM [7]: an attention-based LSTM model which integrates aspect embedding into the sentence input.
- MemNet [6]: a multi-hop memory network with content attention and location attention.
- RAM [28]: a GRU-based approach which uses multiple attention mechanism to capture important information from the sentences.
- BERT [12]: a pre-trained BERT model with stacked Transformer encoder layers to capture the interaction between the aspect and the text.
- MIMN [16]: a multi-hop memory network to capture the interaction of textual and visual modalities.
- ESAFN [18]: an entity-sensitive attention and fusion network to capture the aspect-text and aspect-image dynamics.
- ViLBERT [53]: an extension of BERT with multiple pre-trained Transformer layers over the concatenation of the text and image features extracted from BERT and Faster R-CNN, respectively.
- TomBERT (ResNet) [17]: a target-oriented multimodal BERT architecture, which employs BERT to obtain the aspect-aware text representation, and a widely used image recognition model ResNet [62] to obtain the image representation.

3. It is worth noting that for the task of ACMSC, some comparison systems cannot be adopted because of their model limitations, such as lack of Chinese pre-trained multimodal model (ViLBERT) and aspect-term based text segmentation (ESAFN).

TABLE 4  
Extrinsic Comparisons on Two ATMSC Datasets and One ACMSC Dataset

Modality	Methods	ATMSC				ACMSC	
		TWITTER-2015		TWITTER-2017		ZOL	
		ACC	F1	ACC	F1	ACC	F1
Image	ResNet-Aspect	59.49	47.79	57.86	53.98	44.95	41.68
	Faster R-CNN-Aspect	59.98	37.71	57.94	54.71	56.35	55.12
Text	AE-LSTM [7]	70.30	63.43	61.67	57.97	59.58	58.95
	MemNet [6]	70.11	61.76	64.18	60.80	59.51	58.73
	RAM [28]	70.68	63.05	64.42	61.01	60.18	59.68
	BERT [12]	74.15	68.86	68.15	65.23	62.39	61.55
	MIMN [16]	71.84	65.69	65.88	62.99	61.59	60.51
Text+Image	ESAFN [18]	73.38	67.37	67.83	64.22	—	—
	ViLBERT [53]	73.76	69.85	67.42	64.87	—	—
	TomBERT (ResNet) [17]	76.60±0.40	71.57±0.16	69.42±0.73	67.70±0.50	64.79±0.41	64.27±0.39
	TomBERT (Faster R-CNN)	77.03±0.32	72.85±0.10	69.77±0.33	67.59±0.90	64.71±0.26	64.44±0.38
	<b>HIMT (Ours)</b>	<b>78.14±0.30†</b>	<b>73.68±0.53†</b>	<b>71.14±0.16†</b>	<b>69.16±0.50†</b>	<b>66.83±0.05†</b>	<b>66.58±0.07†</b>
	<b>HIMT w/o Top-<i>N</i> semantics</b>	76.92±0.36	72.62±0.71	69.36±0.28	67.62±0.39	64.80±0.41	64.27±0.39
	<b>HIMT w/o MFT layer</b>	76.52±0.53	72.61±0.30	69.93±0.59	68.49±0.84	64.59±0.43	64.25±0.29
	<b>HIMT w/o Reconstruction Module</b>	76.38±0.44	71.67±0.51	70.18±0.33	68.22±0.66	64.66±0.64	64.41±0.61

† indicates that our full model HIMT is significantly better than TomBERT (ResNet) and TomBERT (Faster R-CNN) with  $p$ -value  $< 0.05$  based on McNemar's significance test. The average results with standard deviations across three runs are reported in the last six rows.

- TomBERT (Faster R-CNN): replacing the image feature extractor in TomBERT with Faster R-CNN.

Table 4 shows the performance of each compared method on three datasets. Note that to better compare HIMT and two variants of TomBERT, we report the average results across three runs in the last six rows. Moreover, we combine the model predictions of all the three runs, and compare HIMT and two variants of TomBERT based on McNemar's significance test, respectively.

In Table 4, we can easily see that our *HIMT* model achieves the best performance on both the ATMSC and ACMSC tasks. Based on the results, we made the following observations: (1) For visual baseline approaches, the method only using hidden representations (i.e., ResNet-Aspect) generally performs the worst, whereas the method with high-level semantic concepts (i.e., Faster R-CNN-Aspect) performs slightly better. (2) Most multimodal approaches generally perform better than their corresponding unimodal baseline approaches. This demonstrates that the image information can well complement the textual information, and thus improve the performance of sentiment classification. (3) Although ViLBERT outperforms most text-only models, it still performs worse than BERT. This is mainly because the ViLBERT model fails to explicitly model the aspect-text and aspect-image interactions, which are crucial for the ABMSA task. (4) Both TomBERT (ResNet) and TomBERT (Faster R-CNN) generally perform better than BERT and ViLBERT on the three datasets, which indicates that incorporating aspect-text and aspect-image interactions can bring additional performance gain. (5) Our HIMT model outperforms the second best approach by 1.63 and 1.86 absolute percentage points on Macro-F1 for the two TWITTER datasets, respectively, and 1.02 absolute percentage points on Weighted-F1 for the ZOL dataset. We verify that the performance improvements are significant. (6) Finally,

removing the Top-*N* semantics, the MFT layer, and the Reconstruction Module respectively will consistently drop the model performance across all the three datasets. This verifies that each of the proposed three components play an indispensable role in our proposed model.

All these observations demonstrate the effectiveness of our HIMT model for aspect-based multimodal sentiment analysis.

### 4.3 Intrinsic Comparisons

To study the impact and the sensitivity of several key components in HIMT, we further conduct intrinsic evaluations.

#### 4.3.1 Top-*N* Semantic Concepts

To explore the sensitivity of Top-*N* semantic concepts, we set the number of kept semantic categories *N* to 0, 1, 3, and 5 for each object, and carry out the experiments respectively.

Based on the results reported in Table 5, we can make the following observations. First, incorporating the Top-*N* semantic concepts generally improves the performance of HIMT without semantic concepts ( $N = 0$ ). Second, the less semantics ( $N = 1$ ) may result in the error propagation issue, while more semantics ( $N = 5$ ) may bring noise. Finally, when *N* is set to 3, we can obtain the best results on both datasets. This indicates that incorporating moderate number of semantic categories for each object (i.e., 3) can alleviate the error propagation and the visual noise issues, and thus leads to better performance.

#### 4.3.2 Multimodal Fusion Transformer Layer

To show the effectiveness of our proposed Multimodal Fusion Transformer (MFT) layer, we propose to remove its components and replace it with a Cross-Modal Transformer (CMT) layer proposed in Tsai *et al.* [43], respectively.

The results are shown in Table 6. First, we can find that both removing the MFT layer and only using aspect-aware image



TABLE 5  
The Result of Keeping Different Top- $N$  Semantic Concepts for Each Object in Our HIMT Model

TOP- $N$ Semantics	TWITTER-2015		TWITTER-2017		ZOL	
	ACC	F1	ACC	F1	ACC	F1
TOP-0	76.92	72.62	69.36	67.62	64.80	64.27
TOP-1	76.38	73.67	69.20	66.63	65.24	64.92
TOP-3	<b>78.14</b>	<b>73.68</b>	<b>71.14</b>	<b>69.16</b>	<b>66.83</b>	<b>66.58</b>
TOP-5	77.14	73.94	70.17	69.26	65.28	64.79

TABLE 6  
The Ablation Study of Our Multimodal Fusion Transformer Layer

Methods	TWITTER-2015		TWITTER-2017		ZOL	
	ACC	F1	ACC	F1	ACC	F1
HIMT	<b>78.14</b>	<b>73.68</b>	<b>71.14</b>	<b>69.16</b>	<b>66.83</b>	<b>66.58</b>
remove MFT	76.76	72.22	70.66	69.64	64.59	64.25
replace MFT w. CMT	77.53	71.73	70.50	68.56	65.52	65.34
using AIR only	59.98	40.61	58.83	54.05	58.03	57.44
using ATR only	77.72	73.07	68.69	67.05	64.08	63.88

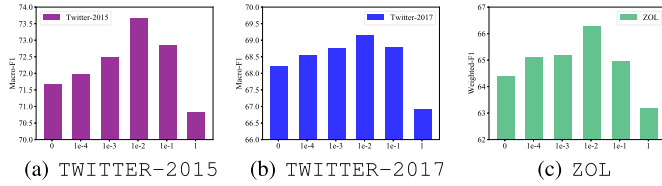


Fig. 4. The sensitivity of the weight of Auxiliary Reconstruction Module.

representations (AIR) or aspect-aware text representations (ATR) will lead to moderate performance drop, which demonstrates the importance of the MFT layer. Second, replacing our MFT layer with the CMT layer results in significant performance drop on both datasets. We conjecture the reason is that the CMT layer may pay too much attention to the inter-modal interactions between text and image representations, but ignore the aspect-text and aspect-image interactions. This further shows that the MFT layer is indispensable to our HIMT model.

#### 4.3.3 Auxiliary Reconstruction Module

To explore the sensitivity of the auxiliary reconstruction module, we report the results of adjusting the weight of the reconstruction loss in our final objective function.

As shown in Fig. 4, neither setting it to an excessive weight (e.g.,  $\lambda = 0.1$ ) or setting it to a small weight (e.g.,  $\lambda = 0$ ) yields the best performance. Across the three datasets,

the changing trend of  $\lambda$  is consistent, and our HIMT model achieves the best performance when  $\lambda$  is set to 0.01. This suggests that it is necessary to incorporate the reconstruction module, but overemphasizing it may lead the model to perform poorly on our main aspect-based multimodal sentiment classification task.

#### 4.3.4 The Sensitivity of the Number of Object Proposals

Finally, we explore the sensitivity of keeping different number of object proposals  $K$  for each image.

In Table 7, we can observe that for all the three datasets, the best performance is obtained when we set  $K = 36$ . A smaller or a larger value of  $K$  will yield the worse performance. This is mainly because a smaller  $K$  may lead our model to ignore the important regions with relatively low detection confidence and a larger  $K$  may bring visual noise to our model. Therefore, the number of object proposals  $K$  in each image is set to 36 in our experiment.

#### 4.4 In-Depth Analysis

In this subsection, we perform in-depth analysis on two Twitter datasets (i.e., TWITTER-2015 and TWITTER-2017) to better understand the advantage and robustness of our HIMT model. Specifically, we first investigate the effects of the size of the training corpus. Second, we compare the computational efficiency of different models. Moreover, we

TABLE 7  
The Sensitivity of the Number of Object Proposals

The number of $K$	TWITTER-2015		TWITTER-2017		ZOL	
	ACC	F1	ACC	F1	ACC	F1
$K = 18$	77.14	73.75	70.34	68.74	65.31	65.27
$K = 27$	77.04	72.48	69.12	67.07	64.72	64.57
$K = 36$	<b>78.14</b>	<b>73.68</b>	<b>71.14</b>	<b>69.16</b>	<b>66.83</b>	<b>66.58</b>
$K = 45$	77.18	72.12	68.56	67.33	65.14	64.86

TABLE 8  
The Performance Comparison of Using Different Sizes of Training Corpora

Size	Methods	TWITTER-2015		TWITTER-2017	
		ACC	F1	ACC	F1
25%	BERT	63.35	50.58	58.83	57.17
	TomBERT (ResNet)	69.43	61.80	62.40	60.60
	TomBERT (Faster R-CNN)	70.68	<b>63.97</b>	63.04	61.74
	HIMT	<b>71.36</b>	63.19	<b>64.58</b>	<b>61.86</b>
50%	BERT	69.62	63.02	61.26	57.74
	TomBERT (ResNet)	73.86	68.58	65.25	63.10
	TomBERT (Faster R-CNN)	74.34	69.55	67.50	64.97
	HIMT	<b>75.41</b>	<b>69.70</b>	<b>68.23</b>	<b>66.54</b>
75%	BERT	72.03	65.74	63.21	59.79
	TomBERT (ResNet)	74.83	70.16	66.36	64.15
	TomBERT (Faster R-CNN)	76.18	71.84	68.39	67.29
	HIMT	<b>77.14</b>	<b>72.86</b>	<b>68.96</b>	<b>67.78</b>
100%	BERT	74.15	68.86	68.15	65.23
	TomBERT (ResNet)	76.60	71.57	69.42	67.70
	TomBERT (Faster R-CNN)	77.03	72.85	69.77	67.59
	HIMT	<b>78.14</b>	<b>73.68</b>	<b>71.14</b>	<b>69.16</b>

select several representative samples to compare the predictions of different methods in Table 10, and visualize the objects with top-ranked attention weights in our AAT layer. Lastly, we manually examine the misclassified samples from HIMT, and summarize them into four categories.

#### 4.4.1 Effects of the Size of Training Corpus

To investigate the robustness of our HIMT model, we conduct experiments with different sizes of training corpora, and compare HIMT with three representative unimodal and multimodal approaches.

First, based on the results shown in Table 8, we can clearly see that as the training data increases, all four approaches explored in this study gradually obtain better performance. Second, it is interesting to observe that no matter how large the training samples we use, HIMT can generally outperform the other three compared systems with a significant margin. Lastly, comparing our HIMT model with the pure text baseline BERT, it is easy to find that their performance gap of using only a small amount of training samples (i.e., 25% and 50%) is much larger than that of using a large amount of training samples (i.e., 75% and 100%), which shows the effectiveness of HIMT in low-resource scenarios. These observations demonstrate the robustness of our proposed approach.

#### 4.4.2 Comparisons of Computational Efficiency

To compare the training and testing efficiency of different methods, we report their training and inference time on one GPU of *NVIDIA TITAN RTX* in Table 9. Note that for fair comparison between TomBERT and HIMT, we use BERT as the text encoder and Faster R-CNN as the image encoder for both methods.

First, during the training stage, since TomBERT and HIMT require more model parameters to encode both the textual and visual information, the two multimodal approaches take much more time than the pure text baseline BERT. Second, HIMT incurs slightly higher computational cost than TomBERT, due to introducing more parameters in Object-Level Semantic Extraction, Multimodal Fusion Transformer Layer, and Auxiliary Reconstruction Module. Moreover, during the testing stage, we find that the inference time of the three approaches is indistinguishable on the two datasets. Based on these observations, we believe that the computational cost of HIMT is generally acceptable considering its performance improvement in Table 4.

#### 4.4.3 Case Study

In Table 10, we show the comparison between predictions of two baseline methods and those of our HIMT model on four test samples.

First, in Table 10a, given the aspect term *officialpepe* as input, our model focuses on the player falling on the ground


TABLE 9  
The Comparison of Training Time (seconds), Inference Time (seconds), and Trainable Parameter Sizes (million) on Different Datasets

Methods	TWITTER-2015			TWITTER-2017		
	Train	Test	#Params	Train	Test	#Params
BERT	291 s	3 s	109 M	343 s	4 s	109 M
TomBERT (Faster R-CNN)	410 s	3 s	235 M	474 s	4 s	235 M
HIMT	488 s	3 s	275 M	507 s	4 s	303 M

The batch sizes for training and testing are set to 32 and 16, respectively.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17, 2024 at 14:22:52 UTC from IEEE Xplore. Restrictions apply.

TABLE 10  
Case Study on Several Test Samples

			
(a) This is @ [officialpepe] <sup>1</sup> . He is a disgrace. He is pathetic. And he is an embarrassment to football. @ [realmadriden] <sup>2</sup>	(b) # [SamHunt] <sup>1</sup> Performs at [Stagecoach] <sup>2</sup> # MusicFestival 2016	(c) Happy Birthday to [Manchester City] <sup>1</sup> hero [Sergio Agüero] <sup>2</sup> , who turns 29 today !	(d) @ [HanrattyDave] <sup>1</sup> And here's a cold, unflinching reminder that [Oscar Isaac] <sup>2</sup> looks like THIS in the movie
Ground Truth: (1-NEG, 2-NEU)	(1-POS, 2-NEU)	(1-NEU, 2-POS)	(1-NEU, 2-NEG)
BERT: (1-NEU✗, 2-NEU✓)	(1-POS✓, 2-NEU✓)	(1-POS✗, 2-POS✓)	(1-NEU✓, 2-NEG✓)
TomBERT: (1-NEU✗, 2-NEU✓)	(1-POS✓, 2-POS✗)	(1-NEU✓, 2-POS✓)	(1-NEU✓, 2-POS✗)
HIMT: (1-NEG✓, 2-NEU✓)	(1-POS✓, 2-NEU✓)	(1-NEU✓, 2-POS✓)	(1-NEU✓, 2-NEG✓)

NEU, POS, and NEG denote neutral, positive, and negative sentiments, respectively. ✓ and ✗ denote the correct and incorrect predictions, respectively. For the aspect terms with incorrect predictions, the objects with top-ranked attention weights in our HIMT model are highlighted by the blue bounding boxes.

which is related to the negative sentiment. It is interesting that BERT made a wrong prediction, probably because the sentiment words and the aspect term are not in the same sentence. In contrast, our model can correct the predictions because of the incorporation of the image features.

Second, in Table 10b, for the aspect term *Stagecoach* with the neutral sentiment, TomBERT made a wrong prediction probably due to the noise from the smiling face in the image, whereas our model mainly focused on the stage equipment such as light and smoke, and thus make a correct prediction.

Moreover, in Table 10c, for the aspect term *Manchester City*, BERT failed to predict its neutral sentiment due to the noise from the positive sentiment words in text. TomBERT and our HIMT model made correct predictions due to the inflow of image representation. Moreover, we can see that our model focused on the background and the suit in the image.

Lastly, in Table 10d, given the aspect term *Oscar Isaac* which is a person name, our model focused on the facial expression and the hand posture in the image, and thus made correct predictions, whereas TomBERT made a wrong prediction due to the irrelevant information from image.

#### 4.4.4 Error Analysis

Furthermore, we also analyze those samples that are misclassified by our HIMT model. Most of the misclassified samples can be grouped into the following four categories:

- Special language phenomenon. For example, given the sentence “@RepAdamSchiff Congrats! You have went full Harry Reid stupid!!” and its aspect term *RepAdamSchiff*, it expresses negative sentiment in the form of irony, and the word *Congrats* affects the prediction of our HIMT model.
- The special form of multi-aspect. In the sentence “#MLB Mercer's double in 12th lifts # Pirates over Rockies 9 – 8,” there are two aspect terms *Pirates* and *Rockies* with Positive and Negative sentiment, respectively. Since the two aspect terms appear in the form of comparison, it is difficult for our model to identify their sentiments.
- Irregular text structure. In the sentence “The women dressed in Calvin Klein Collection. #metgala,” the aspect

term is *metgala*, but it is in an irregular hashtag form in Twitter.

- The mismatch between text and its associated image. In the sentence “Candiens NHL Pet dog jersey shirt (all sizes) NEW Montreal Candiens,” the sentiment of *Candiens* is Neutral, but the visual content is the sign with several letters and colors which is unrelated to the sentence and leads to the wrong prediction.

## 5 CONCLUSION AND FUTURE WORK

In this paper, we first summarized existing studies to ABMSA into two subtasks, i.e., aspect-term based multimodal sentiment classification (ATMSC) and aspect-category based multimodal sentiment classification (ACMSC). Next, we examined the limitations of existing approaches for both tasks of ATMSC and ACMSC, and then proposed a general Hierarchical Interactive Multimodal Transformer (HIMT) model to address these limitations, including incorporating object-level semantics, modeling hierarchical interactions among the input aspect, text, and images, and designing an auxiliary reconstruction module. Experiment results demonstrate that our HIMT model can significantly outperform a number of strong baseline approaches on two benchmark datasets for the ATMSC task and one benchmark dataset for the ACMSC task.

Despite obtaining promising performance, our proposed approach still has several limitations. First, since most existing datasets for both ATMSC and ACMSC tasks are relatively small, our approach bears the risk of over-fitting the training set. Second, our approach assumes that the aspect terms or the aspect categories have been annotated for the input text, which may limit its applications in real-world scenarios. Therefore, we plan to extend our work from the following two perspectives in the future. First, as there are a large amount of unlabeled multimodal posts and product reviews on various social media platforms and E-commerce websites, we plan to follow many existing pre-trained Vision-and-Language models such as LXMERT [52], VL-BERT [65], and UNITER [66] to leverage these unlabeled data to train an effective task-specific pre-training framework, which can potentially improve the performance and robustness of ABMSA approaches. Moreover, another promising direction is to explore the end-to-end

ABMSA task, in which the goal is to simultaneously detect the mentioned aspect terms (or aspect categories) and identify their corresponding sentiments in a joint manner.

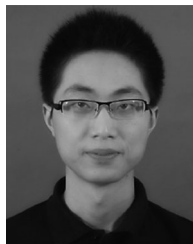
## ACKNOWLEDGMENTS

Jianfei Yu and Kai Chen contributed equally to this paper.

## REFERENCES

- [1] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discov.*, vol. 8, no. 4, 2018, Art. no. e1253.
- [2] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics*, 2011, pp. 151–160.
- [3] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad, "Nrc-Canada-2014: Detecting aspects and sentiment in customer reviews," in *Proc. 8th Int. Workshop Semantic Eval.*, 2014, pp. 437–442.
- [4] D.-T. Vo and Y. Zhang, "Target-dependent twitter sentiment classification with rich automatic features," in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 1347–1353.
- [5] M. Pontiki et al., "SemEval-2016 task 5: Aspect based sentiment analysis," in *Proc. Int. Workshop Semantic Eval.*, 2016, pp. 19–30.
- [6] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2016, pp. 214–224.
- [7] Y. Wang et al., "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2016, pp. 606–615.
- [8] S. Wang, S. Mazumder, B. Liu, M. Zhou, and Y. Chang, "Target-sensitive memory networks for aspect sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 957–967.
- [9] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, and Z. Wu, "Content attention model for aspect based sentiment analysis," in *Proc. World Wide Web Conf.*, 2018, pp. 1023–1032.
- [10] W. Wang, S. J. Pan, and D. Dahlmeier, "Memory networks for fine-grained opinion mining," *Artif. Intell.*, vol. 265, pp. 1–17, 2018.
- [11] Y. Wang, A. Sun, M. Huang, and X. Zhu, "Aspect-level sentiment analysis using as-capsules," in *Proc. World Wide Web Conf.*, 2019, pp. 2033–2044.
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [13] H. Xu, B. Liu, L. Shu, and S. Y. Philip, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 2324–2335.
- [14] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2019, pp. 380–385.
- [15] J. Su et al., "Enhanced aspect-based sentiment analysis models with progressive self-supervised attention learning," *Artif. Intell.*, vol. 296, 2021, Art. no. 103477.
- [16] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, Art. no. 46.
- [17] J. Yu and J. Jiang, "Adapting bert for target-oriented multimodal sentiment classification," in *Proc. Int. Conf. Artif. Intell.*, 2019, pp. 5408–5414.
- [18] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 429–439, Jan. 2020.
- [19] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 513–520.
- [20] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.
- [21] L. Deng and J. Wiebe, "Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2015, pp. 179–189.
- [22] L. Dong et al., "Adaptive recursive neural network for target-dependent Twitter sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 49–54.
- [23] W. Xue and T. Li, "Aspect based sentiment analysis with gated convolutional networks," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2514–2523.
- [24] M. Zhang, Y. Zhang, and D.-T. Vo, "Gated neural networks for targeted sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3087–3093.
- [25] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. Int. Conf. Comput. Linguistics*, 2016, pp. 3298–3307.
- [26] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, Art. no. 721.
- [27] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proc. Int. Conf. Artif. Intell.*, 2017, pp. 4068–4074.
- [28] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 452–461.
- [29] M. Yang, W. Yin, Q. Qu, W. Tu, Y. Shen, and X. Chen, "Neural attentive network for cross-domain aspect-level sentiment classification," *IEEE Trans. Affective Comput.*, vol. 12, no. 3, pp. 761–775, Third Quarter 2021.
- [30] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
- [31] C. Busso et al., "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces*, 2004, pp. 205–211.
- [32] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Commun.*, vol. 53, no. 9/10, pp. 1162–1171, 2011.
- [33] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2015, pp. 2539–2544.
- [34] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.
- [35] P. P. Liang, Z. Liu, A. B. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 150–161.
- [36] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an obviously perfect paper)," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4619–4629.
- [37] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2506–2515.
- [38] P. P. Liang et al., "MultiBench: Multiscale benchmarks for multimodal representation learning," in *Proc. NeurIPS Datasets Benchmarks Track*, 2021, pp. 1–20.
- [39] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.
- [40] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.
- [41] A. Kumar and J. Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4477–4481.
- [42] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.
- [43] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.
- [44] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 223–232.
- [45] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, "Object-based visual sentiment concept analysis and application," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 367–376.
- [46] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.
- [47] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2513–2525, Sep. 2018.

- [48] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7584–7592.
- [49] Q. You, J. Luo, H. Jin, and J. Yang, "Joint visual-textual sentiment analysis with deep neural networks," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 1071–1074.
- [50] Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1008–1017.
- [51] F. Fan, Y. Feng, and D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 3433–3442.
- [52] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proc. Empir. Methods Natural Lang. Process. and the 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 5103–5114.
- [53] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13–23.
- [54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
- [55] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.
- [56] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
- [57] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
- [58] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [59] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5674–5681.
- [60] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1990–1999.
- [61] Y. Cui et al., "Pre-training with whole word masking for chinese bert," 2019, *arXiv:1906.08101*.
- [62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [63] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
- [64] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du, "Analogical reasoning on chinese morphological and semantic relations," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 138–143.
- [65] W. Su et al., "VL-BERT: Pre-training of generic visual-linguistic representations," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–16.
- [66] Y.-C. Chen et al., "UNITER: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 104–120.



**Jianfei Yu** received the BSc and MEng degrees from the Nanjing University of Science and Technology, China, in 2012 and 2015, respectively, and the PhD degree from Singapore Management University, in 2018. He is currently an associate professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include natural language processing, sentiment analysis, information extraction, and question answering.



**Kai Chen** received the BSc degree from the Zhejiang University of Finance and Economics, Hangzhou, China, in 2017, and the MSc degree from the Nanjing University of Science and Technology, Nanjing, China, in 2021. He is currently an algorithm engineer in Baidu, China. His research interests include natural language processing, sentiment analysis, and multimodal learning.



**Rui Xia** received the BSc degree from Southeast University, Nanjing, China, in 2004, the MSc degree from the East China University of Science and Technology, Shanghai, China, in 2007, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, in 2011. He is currently a professor with the School of Computer Science and Engineering, Nanjing University of Science & Technology, China. His research interests include natural language processing, machine learning, and data mining.

► For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/csdl](http://www.computer.org/csdl).