Neurocomputing 455 (2021) 47–58

Contents lists available at ScienceDirect

Neurocomputing

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

MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis

Jie Zhou [a,*], Jiabao Zhao [a], Jimmy Xiangji Huang [b], Qinmin Vivian Hu [c], Liang He [a]

[a] School of Computer Science and Technology, East China Normal University, Shanghai 200241, China
[b] Information Retrieval and Knowledge Management Research Lab, York University, Toronto, Ontario M3J 1P3, Canada
[c] The School of Computer Science, Ryerson University, Toronto, Ontario M5B 2K3, Canada

ARTICLE INFO

Article history:
Received 1 May 2020
Revised 12 May 2021
Accepted 13 May 2021
Available online 18 May 2021
Communicated by Zidong Wang

Keywords:
Sentiment analysis
Multimodal
Aspect-based sentiment analysis
Deep learning

ABSTRACT

Aspect-based sentiment analysis has obtained great success in recent years. Most of the existing work focuses on determining the sentiment polarity of the given aspect according to the given text, while little attention has been paid to the visual information as well as multimodality content for aspect-based sentiment analysis. Multimodal content is becoming increasingly popular in mainstream online social platforms and can help better extract user sentiments toward a given aspect. There are only few studies focusing on this new task: Multimodal Aspect-based Sentiment Analysis (**MASA**), which performs aspect-based sentiment analysis by integrating both texts and images. In this paper, we propose a mutimodal interaction model for MASA to learn the relationship among the text, image and aspect via interaction layers and adversarial training. Additionally, we build a new large-scale dataset for this task, named **MASAD**, which involves seven domains and 57 aspect categories with 38 k image-text pairs. Extensive experiments have been conducted on the proposed dataset to provide several baselines for this task. Though our models obtain significant improvement for this task, empirical results show that MASA is more challenging than textual aspect-based sentiment analysis, which indicates that MASA remains a challenging open problem and requires further efforts.

1. Introduction and motivation

Nowadays, as a major social media platform for expressing experiences and sharing the opinion about service, products, and travel, the Internet provides extensive contents of users' opinion and sentiment about rich topics [1]. This information is expressed in multiple formats, such as reviews, tags, browser behavior, and shared media objects. The analysis of such information plays an essential role in the area of opinion mining, affective computing, and sentiment analysis. It can predict human decision making and enables some applications like public opinion analysis, brand monitoring, and political voting forecast [2]. As a fundamental subtask of sentiment analysis, aspect-based sentiment analysis provides valuable insights to both consumers and businesses. It helps companies to measure satisfaction and improve their products or services [3,4]. So far, the computational analysis of aspect-based sentiment mostly concentrates on opinionated texts (e.g., comments, tweets, and reviews) [3,5]. However, limited efforts have been conducted to analyze sentiments from visual content such as images, which is becoming a key media type on the web. Recently, social media users are increasingly using additional images to express their experiences and opinion. Such rich-source textual and visual content can help better extract user sentiments toward different aspects.

Motivated by the needs to use large-scale social multimodal content for aspect-based sentiment analysis, we focus on the new task Multimodal Aspect-based Sentiment Analysis (**MASA**), where the input is in the form of an image plus some text that describes the input image. MASA consists of two subtasks: aspect extraction (**AE**) and aspect polarity prediction (**AP**). The goal of AE is to extract the aspect of the samples via the image-text pair and the AP aims to predict the sentiment polarity with respect to the aspect. Taking Fig. 1 as an example, in the above one, two samples both contain a car and the sentiment polarities of them are negative and positive since the left car is crushed while the right car is new and cool with a clear sky. In the blow one, the sentiments of the aspect "dog" are positive and negative, respectively. We first need to extract the aspects (e.g., car, dog and so on) from the image-text pair and then we need to determine their sentiment polarities (e.g., positive or negative) according to the image and text.

* Corresponding author.
E-mail address: jzhou@ica.stc.sh.cn (J. Zhou).

https://doi.org/10.1016/j.neucom.2021.05.040
0925-2312/

Fig. 1. An example of multimodal aspect-based sentiment analysis. This task includes two subtasks: Aspect Extraction (**AE**) and Aspect Polarity Prediction (**AP**). First, we extract the aspect (e.g., car or dog) in the image-text pair. Then, we judge the sentiment (e.g., positive or negative) with respect to the extracted aspect.

To better explore the MASA task, we build and release a new large-scale dataset for multimodal aspect-level sentiment analysis, named **MASAD**. Table 1 shows the comparison of the public aspect-level sentiment analysis with our proposed MASAD. One problem of these existing datasets is that most of these datasets focus on the textual information, while the visual information is ignored. Another problem is that the limited number of domains does not guarantee the generalization capability required in real applications. Besides, the limited number of data size largely limits the performance of this task. To address previous drawbacks, we propose this MASAD dataset, which consists of 38 k samples within seven domains (e.g., Food, Goods, Buildings, Animal, Human, Plant, and Scenery) and 57 aspects. Our dataset is larger than the existing available datasets of this field and is the only one that contains both text and image for this task.

Moreover, we propose a multimodal interaction model to assess the challenges of MASA and provide strong baselines for this task. To be specific, we first use image and text encoders to learn the representation of the image-based and text-based on state-of-the-art visual and textual technology. Then we design a multi-modal interaction layer to learn the relationships among the aspect, image, and text. Effectively fusing this diverse textual and visual information is non-trivial and poses several challenges to the underlying problem. Our model combines the strengths of the text and the image representations by extracting interactive information between them. Furthermore, we design an adversarial training strategy to align text and image features into one common space. A series of experiments also show the great advantages of our models.

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

- We propose a multimodal interaction model for the new multi-modal aspect-based sentiment analysis (**MASA**) task. Different from the existing aspect-based sentiment analysis task, which judges the sentiment polarity of the aspect based on textual information, this new task infers the sentiment for the given aspect based on both texts and images. Our model can learn the interactive relation among the aspect, text, and image effectively.
- Since the multimodal data (e.g., text and image) is not in a common space, we design an adversarial training to align the feature representations of image and text into a shared space.
- We present and release a large-scale dataset for multimodal aspect-based sentiment analysis, namely **MASAD**, which involves 57 categories in seven domains and contains 38 k samples with image-text pair. This dataset provides a new perspective for aspect-based sentiment classification and a new benchmark for MASA.
- The extensive experimental results indicate the great advantage of our model. However, detailed analysis on the experiments shows that MASA is more challenging than textual aspect-based sentiment analysis and intensive efforts are needed to improve the performance.

Table 1
The comparison of the public aspect-level sentiment analysis.

	Year	#Samples	#Aspects	#Domain	Text	Image
$\text{Mitchell}_{\text{EN}}$ [6]	2013	10,000	10	1	✓	×
SemEval-2014 [5]	2014	7,686	5	2	✓	×
SemEval-2015 [7]	2015	4,766	26	3	✓	×
$\text{SemEval-2016}_{\text{EN}}$ [8]	2016	5,984	28	2	✓	×
SentiHood [9]	2016	5,215	2	1	✓	×
Multi-ZOL [10]	2019	5,228	40	1	✓	✓
Twitter-15/17 [11]	2019	3,179/3,562	–	1	✓	✓
MASAD (our)	2019	38,532	57	7	✓	✓

2. Related work

We divide the relevant work into two groups: aspect-based sentiment analysis and multimodal sentiment analysis. There are a large number of studies on the topic of aspect-based sentiment classification [4,5,7,8,12] and multimodal sentiment analysis [2,13]. Here we mainly review the work that is most related to our research.

2.1. Aspect-based sentiment analysis

Aspect-based sentiment analysis plays a significant role in sentiment analysis [3,14]. Early work mainly used traditional machine learning algorithms, which highly depended on the quality of extensive hand-craft features [15,16]. With the advances of neural networks, various deep learning models are of growing interest in aspect-based sentiment analysis for their ability to learn the representation of text automatically [17–24]. Attention mechanisms [25] have been adopted to capture the important parts with respect to the given aspect [26–31]. Sentiment commonsense knowledge are integrated into this task to improve the performance [32,33]. In addition, pre-trained models (such as BERT, ELMo) are utilized for aspect-based sentiment analysis [34,35].

In recent years, aspect-based sentiment analysis task has been organized by some workshops and conferences [5,7,8]. These competitions provided training datasets and the opportunity for a comparison of different methods on the test set. Dong et al. [36] presented a manually annotated dataset for target-dependent twitter sentiment analysis. And Fan et al. [37] labeled the opinion words with respect to the given aspect for these datasets. SentiHood [9] was a benchmark dataset that was annotated for the targeted aspect-based sentiment analysis task in the domain of urban neighborhoods. Michell et al. [6] released a dataset[1] that includes about 30 k Spanish tweets and 10 k English tweets labeled for named entities with the sentiments. Recently, Jiang et al. [38] proposed a new datasets, where the samples contain more than one aspects with different sentiments.

Despite these advances of methods and datasets in aspect-based sentiment analysis, almost all of them focus on how to perform classification based on the textual context, while the visual information as well as multimodal content is ignored. More recently, Xu et al. and Yu et al. [10,11] proposed the task of multimodal aspect-level sentiment analysis. Different from their researches, we aim to explore the effectiveness of the interaction among the aspect, text and image in this paper. We propose a multimodal interaction model for this new task. We also present and release a large-scale dataset for this new task. The dataset we provided facilitates studying MASA relative to conceptually different and diverse aspects.

[1] http://www.m-mitchell.com/code/index.html.

2.2. Multimodal sentiment analysis

Multimodal sentiment analysis has recently attached attention due to the tremendous growth of many social media platforms such as Twitter, Flickr, Facebook, and so on [39,40]. It depends on the information obtained from more than one modality (i.e., text and image) for the analysis. Multimodal sentiment analysis has become a very challenging problem for the researchers [41]. A survey of the literature showed that multimodal sentiment analysis is relatively a new area compared to textual sentiment analysis [40]. Ghosal et al. [42] developed an RNN-based multimodal attention model to utilize the contextual information for utterance-level sentiment analysis. Poria et al. [13] introduced an LSTM-based method that leveraged the contextual information to capture the inter-dependencies between the utterances. In another work, a user opinion based model is proposed to combine the three modality inputs (i.e., text, visual and acoustic) by applying a multi-kernel learning based method [40]. Zadeh et al. [43] developed multi-attention blocks (MAB) to capture information across text, visual and acoustic.

However, almost all of them focus on the multimodal sentiment analysis for the whole sample rather than aspect level. Xu et al. [10] integrated the image information into the attention mechanism for aspect-based multi-modal sentiment analysis task. However, the interaction and alignment between the text and image information are ignored by this work. In this paper, we propose to learn the interaction between the aspect and multi-modal, and the interaction among multi-modal information. Moreover, we design an adversarial training strategy to align the text and image representations into the same space.

3. Task description

In this paper, we focus on the new task, namely multimodal aspect-level sentiment analysis (MASA), which aims to perform the aspect-based sentiment analysis [3,5] based the textual and visual information. This task consists of two subtasks: Aspect Extraction (**AE**) and Aspect Polarity Prediction (**AP**). The goal of **AE** is to extract the aspect (e.g., dog, car) in the text-image pair. Then **AP** aims to judge the sentiment polarity for the given aspect via the text-image pair. Formally, the definitions of these two subtasks are given as follows.

Aspect Extraction (**AE**): Given a predefined set of aspects A and a text-image pair $P(T, I)$, the aim of this task is to identify the aspect $a \in A$ expressed in the text-image pair $P(T, I)$. The number of the aspects is $|A|$. The text $T = \{w_1, w_2, \ldots, w_n\}$ consists of n words. For example, in Fig. 1, the above two samples contain aspect "car" and the below samples contain aspect "dog".

Aspect Polarity Prediction (**AP**): For this subtask, aspect a for each text and image pair $P(T, I)$ is provided. The goal is to determine the sentiment polarity $c \in \{N, P\}$ of the aspect a discussed in each text and image pair, where N and P denote the "negative" and "positive" sentiment polarities respectively. For instance, the

user expresses negative and positive sentiments over aspect "car" in the above samples, respectively (Fig. 1). And the sentiments for aspect "dog" in the below samples are positive and negative, respectively.

4. Dataset

4.1. Data collection

We collect and label our dataset based on the publicly available Visual Sentiment Ontology (VSO) dataset[2] [44] and Multilingual Visual Sentiment Ontology (MVSO) dataset[3] [2], which are the largest available datasets for visual sentiment analysis. VSO dataset was collected from Flickr.[4] We select Flickr since there is some existing work of multimedia research using it [44,2], and to be specific, Jin et al. [45] presented that Flickr has two advantages popularity and availability for utilizing the "wisdom of the social multimedia". However, the VSO dataset does not provide the text and sentiment polarities towards the given objects. Moreover, it contains a lot of noise and many images in it have no clear sentiments. We select the samples which can express obvious sentiments (about 38 k samples) from parts of VSO dataset (about 120 k samples) and summarize them into seven domains. Then we crawl the descriptions (text information) of images and clean the data for each aspect to ensure the high-quality of each sample with image-text pair. To obtain the given aspect's sentiment polarity, we label the datasets through crowdsourcing with three annotators. In particular, we ask three workers to label the results via Ali Crowdsourcing Platform.[5] We present the image-text pair to the annotators and ask them to label the sentiment polarities towards the given aspect. Then, we obtain the label through voting. We calculate the Krippendorff's alpha coefficient [46] to measure the inter-annotator agreement of the manual annotation. The value is 0.850, which indicates the high agreement of the labeled data. We updated the dataset in github with a readme.[6]

4.2. Dataset description

With the widespread application of aspect-based sentiment analysis in e-commerce, the datasets provided by SemEval [5,7,8] are widely used in this field. Most of the existing aspect-based sentiment analysis datasets only contain text, while the visual information is ignored. To address this problem, we present and release a large-scale multimodal aspect-based sentiment analysis dataset (MASAD), which contains both texts and images. The main statistics of our proposed MASAD is provided in Table 2. This dataset consists of seven domains, including food, goods, buildings, animal, human, plant, and scenery. Each domain contains several aspects with positive and negative samples. There are a total of 57 predefined aspects, such as dog, cat, car, etc. As shown in Fig. 2, several typical examples of two aspects "Leaves" and "Face" are given. Notably, all the samples have an obvious sentiment polarity and the image and text are both critical for the classification. In particular, our dataset consists of 38,532 text and image pairs. The number of samples with positive and negative sentiments is 23,429 and 15,103, respectively. The dataset is split into two sets: training and testing. We have 29,588 and 8,944 samples for its corresponding set.

[2] https://visual-sentiment-ontology.appspot.com/
[3] http://mvso.cs.columbia.edu/
[4] https://www.flickr.com/.
[5] https://newjob.taobao.com/#!/.
[6] https://github.com/12190143/MASAD.

5. Methods

In this section, we aim to find a straightforward architecture that provides good performance for the MASA task. In particular, we propose a multimodal aspect extraction (**MMAE**) method and a multimodal aspect polarity prediction (**MMAP**) method for aspect extraction (AE) and aspect polarity prediction (AP) respectively. The frameworks of these models are shown in Fig. 3. We first use text encoder and image encoder to extract the text and image representation via the state-of-the-art textual and visual methods. We also propose an aspect embedding the obtain the representation of the aspect. To learn the relationships among the text, image, and aspect, we design three interaction mechanisms to learn the interaction between the text and image, the text and aspect, and the image and aspect. Additionally, we design an adversarial training strategy to align the feature representation of text and image into a common space. The details are showing in the following sections.

5.1. Multimodal aspect extraction (MMAE)

In this section, we propose our multimodal aspect extraction (MMAE) method for AE. Given the text-image pair $\{T, I\}$, the goal of this task is to identify the aspect expressed in it. The framework of the MMAE model is shown in Fig. 3. Our model consists of four parts: Image Encoder, Text Encoder, Multimodal Interaction Layer, and Aspect Classification. Image Encoder and Text Encoder are used to extract the feature representations of image and text. Then we develop a multimodal interaction layer to learn the interaction between the text and image. Finally, the text-image pair representation is fed to an aspect classification layer. Details are presented as follows.

Text Encoder Each word w_i in the text is mapped into a low-dimensional continuous vector space $x_i \in \mathbb{R}^{d_w}$, which is calculated by looking up the word embedding $E_w \in \mathbb{R}^{d_w \times |V|}$. Here d_w denotes the dimension of the word embedding and $|V|$ is the size of vocabulary. We use the pre-trained 300-dimensional GloVe [47] embedding to represent words.

Bi-directional long short-term memory (Bi-LSTM) [48] model is employed to accumulate the context information from word embedding. The Bi-LSTM contains a forward $\overrightarrow{LSTM}$ and a backward $\overleftarrow{LSTM}$ which read the text from w_1 to w_n and w_n to w_1 respectively. The contextualized representation for each word is computed as follows:

$$\overrightarrow{h_i} = \overrightarrow{LSTM}(x_i), i \in [1, n] \tag{1}$$

$$\overleftarrow{h_i} = \overleftarrow{LSTM}(x_i), i \in [n, 1] \tag{2}$$

Finally, the last hidden state of the Bi-LSTM is used as the feature representation of the text R^T:

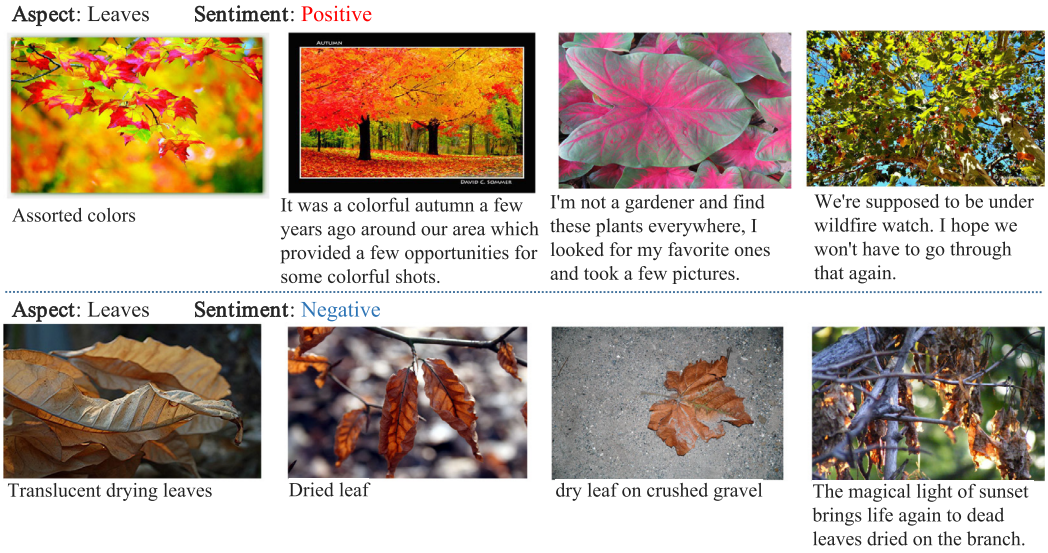
$$R^T = \left[\overrightarrow{h_n}; \overleftarrow{h_n}\right] \tag{3}$$

where $R^T \in \mathbb{R}^{d_T}$ is obtained by concatenating the hidden states $\overrightarrow{h_i} \in \mathbb{R}^{d_h}$ and $\overleftarrow{h_i} \in \mathbb{R}^{d_h}$. And ; donates the concatenate operator, d_h is the dimension of the hidden states and $d_T = 2d_h$ is the dimension of the text features.

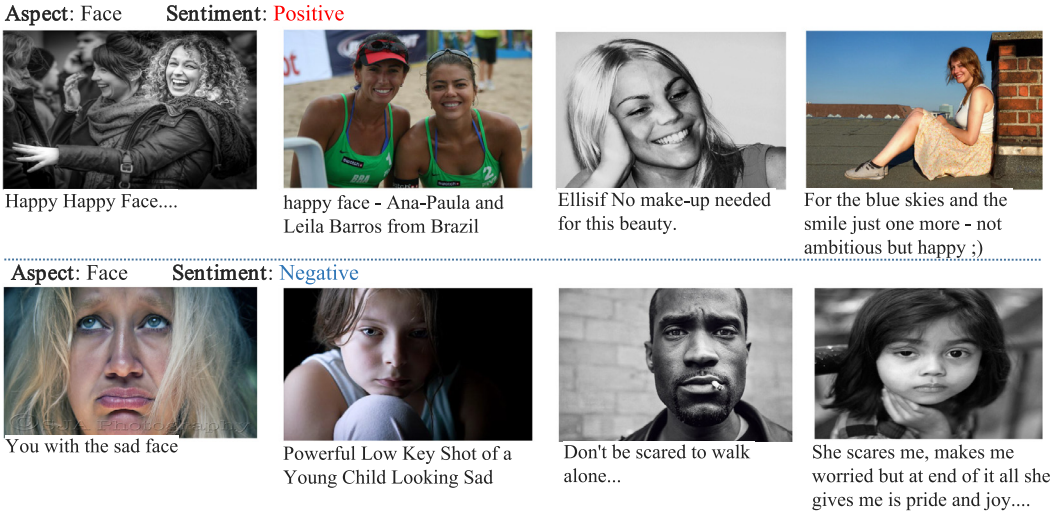
Image Encoder In this part, we adopt an Image Encoder to learn the feature representation of the image. Recently, the residual network (ResNet) [49] achieved great success on a wide range of vision tasks. Gajarla and Gupta [50] applied pre-trained ResNet for visual sentiment analysis and achieved good performance. More specifically, we adopt the ResNet-50 [49] as the image enco-

Table 2
Statistical information of our proposed MASAD dataset.

	Train			Test			Total		
	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Food	2360	433	2793	592	109	701	2952	542	3494
Goods	2671	1674	4345	743	512	1255	3414	2186	5600
Buildings	1450	970	2420	367	245	612	1817	1215	3032
Animal	3023	2208	5231	1126	670	1796	4149	2878	7027
Human	1999	1838	3837	503	464	967	2502	2302	4804
Plant	2819	2607	5426	1269	947	2216	4088	3554	7642
Scenery	3600	1936	5536	907	490	1397	4507	2426	6933
Total	17922	11666	29588	5507	3437	8944	23429	15103	38532



(a) Examples of aspect "Leaves" with different sentiments.



(b) Examples of aspect "Face" with different sentiments.

Fig. 2. Several examples of two typical aspects with different sentiment polarities in our MASAD dataset.

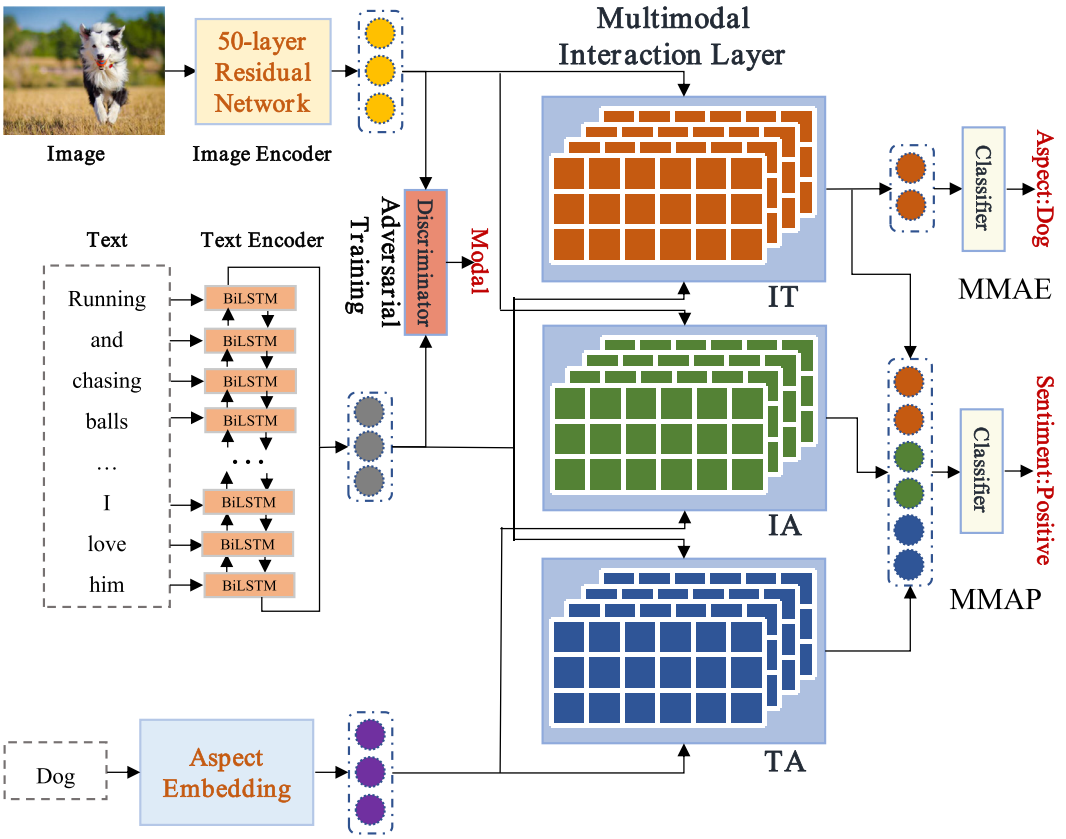


Fig. 3. The framework of Multimodal Aspect Extraction (**MMAE**) and Multimodal Aspect Polarity Polarity (**MMAP**). For MMAE, we first learn the representation of the image and text via image and text encoders. To align the representations of the text and image in the same space, we design an adversarial training strategy by predicting the modal based on the input (text representation or image representation). Then, we design a multimodal interaction layer to learn the relationships between the text and image. Finally, the interactive representation of the image-text pair is fed to aspect classifier for predicting the aspect. Similar to **MMAE**, for **MMAP**, we utilize image and text encoders to extract the representation of the image and text respectively. In addition, we design an aspect embedding to learn the representation of the aspect. Then, we adopt three interaction parts to learning the relationships between the image and aspect, the aspect and text, and the image and text. Finally, we concatenate these three interactive representations and feed them into the sentiment classifier to predict the sentiment for the given aspect..

der to extract the feature representation of the image. The ResNet-50 is a 50 layer Residual Network, which obtains excellent results by extremely deep residual nets. Note that we use the parameters pre-trained on the ImageNet[7] as the initialization parameters and fine-tune the model on our dataset. Finally, the feature representation of the image I can be calculated as:

$$R^I = \text{ImageEncoder}(I) \tag{4}$$

where $R^I \in \mathbb{R}^{d_I}$ is the feature representation of the image, d_I is the dimension of the image features and ImageEncoder is a ResNet-50 model here.

Multimodal Interaction Layer The main challenge in multimodal sentiment analysis lies in properly using the information extracted from multiple modalities, such as text and image. Although it is often argued that the modalities are always beneficial for enhanced performance, it is notable that not the relationships among the modalities are vital. To obtain a better representation of the text-image pair, we design a multimodal interaction layer to learn the inter-modality relationship between the text and image. As shown in Fig. 3, we propose an Inter_{IT} to learn the interaction between the image and text.

$$R^{IT} = \text{Inter}_{IT} = R^I \times A \times R^T + b \tag{5}$$

where $A \in \mathbb{R}^{d_T \times d_{IT} \times d_I}$ is the learnable weights and b is the bias. $R^{IT} \in \mathbb{R}^{d_{\text{IT}}}$ is the interactive feature representation of text-image pair and d_{IT} is dimension of R^{IT}.

Aspect Classification The final step is to feed the representation of text-image pair R_{IT} to a multilayer perceptron (MLP) and softmax layer for aspect distribution prediction:

$$p^{AE} = \text{Softmax}\left(W_{AE}R^{IT} + b_{AE}\right) \tag{6}$$

where $W_{AE} \in \mathbb{R}^{d_h \times |A|}$ and $b_{AE} \in \mathbb{R}^{|A|}$ are the learnable parameters, A is the number of aspect classes.

The loss function is defined by the cross-entropy of the predicted and true label distributions for training:

$$\mathbb{L}_{AE} = -\frac{1}{N}\sum_{i}^{N} \log(p_a^{AE}) \tag{7}$$

where N is the number of instances in the dataset, a is the true class of the i_{th} instance, and p_k^{AE} indicates the k^{th} value of the vector p^{AE}.

5.2. Adversarial training for alignment

Although interaction representation of text and image are learned by the multimodal interaction layer, there is no guarantee that features of image and text are in the same space. Inspired by [51], we utilize the adversarial training to align the space of image and text into the same space (Fig. 3). In particular, we adopt a dis-

[7] https://pytorch.org/docs/stable/torchvision/models.html.

criminator to judge the representation comes from text (y^T) or image (y^I). It minimizes the cross-entropy of the predicted label distribution $p(y_R|R)$ and the true label (y^T or y^I):

$$\min_{\theta^d} \mathscr{L}_d\left(\theta^d\right) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(p\left(y_R = y^I|R^I\right)\right) + \log\left(p\left(y_R = y^T|R^T\right)\right) \tag{8}$$

where y^I and y^T represent the label of the feature representation belong to image or text, respectively. θ^d is the parameters of discriminator. Specifically, we use an MLP and a softmax layer as the modal discriminator.

$$p(y_R|R) = \text{Softmax}(W_d R + b_d) \tag{9}$$

where R is the vector representations of the text/image (R^T/R^I) and $\theta^d = \{W_d, b_d\}$.

Our generator is quite different from the traditional generator in multi-criteria tasks [51]. To be specific, the traditional generator always plays against the discriminator to obtain the invariant representations of text and image that the discriminator can not distinguish. In particular, the generator (image encoder and text encoder) plays an adversarial game with the discriminator, making it difficult to discriminate the feature label (text (y^T) or image (y^I)). Thus, our generator aims to maximize the feature label prediction of the generated feature according to text or image as follows:

$$\max_{\theta^g} \mathscr{L}_g(\theta^g) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(p\left(y_R = y^I|R^I\right)\right) + \log\left(p\left(y_R = y^T|R^T\right)\right) \tag{10}$$

where θ^g is the parameters of the generator (e.g., image encoder and text encoder).

5.3. Multimodal aspect polarity prediction (MMAP)

We develop a multimodal aspect polarity prediction (MMAP) model for task AP, which aims to determine the sentiment polarity of the given aspect in the text-aspect pair. Fig. 3 shows the framework of our MMAP model. The same as MMAE, we use Image Encoder and Text Encoder to obtain the image's feature representation R^I via Eq. 3 and text's feature representation R^T via Eq. 4 respectively. Also, we design an Aspect Embedding to learn the representation of the given aspect. For the Interaction Layer, we propose three strategies to model the interaction between image and aspect, image and text, and aspect and text to capture the relationships between them more effectively. We give details in the following sections.

Aspect Representation To make full use of aspect information, we propose to learn an embedding vector for each aspect. Inspired by [26], we obtain the corresponding aspect representation R^A by looking up an aspect embedding matrix $E_A \in \mathbb{R}^{d_{\mathbb{A}}}$, which is initialized by GloVe, and updated during the training process. d_A is the dimension of the aspect embedding.

Multimodal Interaction Layer How to model the relationships among the text, image, and aspect plays a significant role in this task. Thus, we design our multimodal interaction layer, which consists of three parts, namely $\text{Inter}_{IA}, \text{Inter}_{IT}, \text{Inter}_{AT}$, which model the interaction between image and aspect, image and text, and aspect and text, respectively.

$$R^{IA} = \text{Inter}_{IA} = R^I \times A_{IA} \times R^A + b_{IA} \tag{11}$$

$$R^{IT} = \text{Inter}_{IT} = R^I \times A_{IT} \times R^T + b_{IT} \tag{12}$$

$$R^{AT} = \text{Inter}_{AT} = R^A \times A_{AT} \times R^T + b_{AT} \tag{13}$$

where $A_{IA} \in \mathbb{R}^{d_T \times d_{IA} \times d_A}$, $A_{IT} \in \mathbb{R}^{d_I \times d_{IT} \times d_T}$ and $A_{AT} \in \mathbb{R}^{d_A \times d_{AT} \times d_T}$ are the learnable weights and b_{IA}, b_{IT} and b_{AT} are the bias. $R^{IA}_{\mathbb{IA}}, R^{IT}_{\mathbb{IT}}$ and $R^{AT}_{\mathbb{AT}}$ are the interactive feature representation of image and aspect, image and text, and aspect and text respectively. And d_{IA}, d_{IT} and d_{AT} are the dimension of R^{IA}, R^{IT} and R^{AT} respectively.

The final representation of the text-image pair towards the aspect is calculated as follows:

$$R^{ITA} = \left[R^{IT}; R^{IA}; R^{AT}\right] \tag{14}$$

where ; denotes the concatenate operator, $R^{ITA}\mathbb{R}^{d_{IT}+d_{IA}+d_{AT}}$ indicates the final representation of the aspect in the text-image pair.

Sentiment Classification The final step is to feed R_{ITA} to a softmax layer for sentiment distribution prediction:

$$p^{AP} = \text{Softmax}\left(W_{AP}R^{ITA} + b_{AP}\right) \tag{15}$$

where $W_{AP} \in \mathbb{R}^{d_h \times |C|}$ and $b_{AP} \in \mathbb{R}^{|C|}$ are the learnable parameters, C is the number of sentiment classes.

The loss function is defined by the cross-entropy of the predicted and true label distributions for training:

$$\mathscr{L}_{AP} = -\frac{1}{N}\sum_{i}^{N} log(p_c^{AP}) \tag{16}$$

where N is the number of instances in the dataset, c is the true class of the i_{th} instance, and p_k^{AP} indicates the k^{th} value of the vector p^{AP}.

5.4. Joint training

Finally, we combine the our two subtask and adversarial objective functions for joint training. For aspect extraction, the final loss function can be calculated as follows:

$$\mathscr{L} = \mathscr{L}_{AE} + \mathscr{L}_d\left(\theta^d\right) - \mathscr{L}_g(\theta^g) \tag{17}$$

For aspect polarity prediction, the final loss function can be computed as follows:

$$\mathscr{L} = \mathscr{L}_{AP} + \mathscr{L}_d\left(\theta^d\right) - \mathscr{L}_g(\theta^g) \tag{18}$$

6. Experimental results and analysis

In this section, we conduct extensive experimental results on two subtasks, AE and AP on our MASAD dataset to evaluate our proposed methods and provide several baselines for this task. We first present the implementation details of our models. Then we discuss the experimental results of AE and AP. To be specific, we compare the results using text only, visual only, and their combination over seven domains of MASAD with various metrics. In addition, to verify the effectiveness of our model, we compare our interaction model with concatenation, which simply concatenates the representations of the image, text, and aspect. To make the abbreviations easier to understand, we list the specific nouns with the corresponding abbreviations (Table 3). These results also provide extensive benchmarks for this task.

6.1. Implementation details

In our experiments, word embedding vectors and aspect embedding vectors are initialized with 300-dimension GloVe [47] vectors and fine-tuned during the training, the same as [52]. The dimension of hidden state vectors d_h, and image feature d_I are 300. Words out of vocabulary GloVe, position embedding and

Table 3
The abbreviation of the specific noun.

Noun	Abbreviation
Aspect Extraction	AE
Aspect Polarity Prediction	AP
Multimodal Aspect Extraction	MMAE
Multimodal Aspect Polarity Predition	MMAP
Multimodal Asepct-based Sentiment Analysis	MASA
Multimodal Asepct-based Sentiment Analysis Dataset	MASAD

weight matrices are initialized with the uniform distribution $U(-0.1, 0.1)$, and the biases are initialized to zero. Adam [53] is adopted as the optimizer with a learning rate of 0.001 and min-batch size 64. We implement our neural networks with Pytorch.[8] We split 10% from training data as the development set and keep the optimal parameters based on the best performance on the development set. We adopt Accuracy (Acc.) and Macro-Average F1 (F1) to evaluate the model performance, which are the primary metrics used in aspect-based sentiment analysis [28,18].

6.2. Main results

To verify the effectiveness of our model, we compare our model with several strong baselines. We split the baselines into two parts: textual and multi-modality aspect-based sentiment analysis. The details are shown as follows.

First, we will introduce some textual aspect-based sentiment analysis baselines, which are widely used in this task.

- **ATAE-LSTM** takes aspect information into account via attention mechanism to capture the salient parts of the sentence [26].
- **IAN** designs an interaction attention mechanism to model the relationships between the aspect and sentence [27].
- **RAM** proposes a weighted memory mechanism to capture the sentiment information of the given aspect [28].
- **TNet** employs a CNN layer to extract important features from the transformed word representations originated from a Bi-RNN layer [54].

Since all the above methods are based on text, we also provide some methods about multimodal aspect-based sentiment analysis.

- **MIMN** adopts two interactive memory networks to model the textual and visual information with the given aspect by learning the interaction between cross-modality data [10].
- **TomBERT** proposes a multimodal BERT model to obtain aspect-aware textual representations via multimodal interaction [11].

Different from the existing models, we first learn the inter- and intra- interaction to learn the relationships among the aspect, text and image. Then, to map the image and text representation into a common space, we design an adversarial training strategy.

Table 5 reports the experimental results of our method and the baselines. From this table, we can obtain the following observations. First, multimodal based models perform better than text-based models. To be specific, MIMN and TomBERT perform better than all the textual aspect-based baselines. Second, our multi-modal aspect-based sentiment classification model performs better than the state-of-the-art methods (e.g., TomBERT). It indicates that our model can learn the interaction between aspect, text and image well. In addition, our adversarial training strategy can align the representation of image and text into a common space effectively.

For Food domain, the accuracy is relatively higher than the Macro-F1 score of the baselines (about 5 points) since the distribution of samples in the this domain is unbalanced. Moreover, integrating multimodal information into aspect-based sentiment information can enhance the sample representations. The interaction and alignment among multimodal information can further improve the performance. The difference between accuracy and Marco-F1 is less than 3 points for our MMAP model, which indicates that our model can reduce the influence of unbalance, to a certain extent.

6.3. Performance of aspect extraction (AE)

We present the results of aspect extraction on our dataset in this section. Table 4 shows the accuracy and F1 of aspect extraction over the whole dataset. In addition, Table 6 reports the experimental results over seven domains with text only, image only, and their combination. From these tables, we find the following observations.

- The model with text performs much better than the one with an image. We find that the differences between some aspects in images are relatively similar, which largely limits the performance of the model. It is worth exploring how to extract the aspect from the image for this task more effectively.
- The model with the interaction layer outperforms the model with concatenation, which indicates that our interaction layer can capture the relationships between the text and image more effectively.
- However, we find that simply concatenating the representation of the image and text sometimes even performs worse than the single-modal model (text-based models). It denotes that designing an effective interaction model is important for this task.
- Our adversarial training can significantly improve performance, which indicates that this module can align the representation of the text and image effectively.
- Moreover, we find that our model obtains 70.32% in terms of accuracy with 57 classes, which indicates the high quality of our MASAD dataset.

6.4. Performance of aspect polarity prediction (AP)

In this section, we present the results of the aspect polarity prediction (AP) task. Table 7 shows the experimental results of AP over the whole dataset. Table 8 reports the Accuracy and F1 of AP with text only, image only and their combination respectively. For this task, we need to model the relationships among the aspect, text, and image. From the experimental results, we observe that the model with our multimodal interaction layer outperforms the one with the text only, image only, and combination of them, which verify the great advantage of our interaction model. Similarly, the text-based models perform better than image-based models for this task. Inferring the sentiment polarity of the image is more challenging than the text. It is observed that our MMAP model obtain 90.14% in terms of F1, which indicates the high qual-

Table 4
The experimental results of aspect extraction over the whole dataset. "(w/o) Adv" means the corresponding model without adversarial training.

		Acc.	F1
Text	BiLSTM	62.43	56.72
Image	ResNet50	30.43	17.39
MMAE	Concatenation	64.07	56.28
	Interaction	**70.32**	**68.92**
MMAE (w/o Adv)	Concatenation	62.11	55.32
	Interaction	69.23	68.42

Table 5
The experimental results of subtask aspect polarity prediction over seven domains over our model and strong baselines.

	Text								Multimodal					
	ATAE-LSTM		IAN		RAM		TNet		MIMN		TomBERT		MMAP (Ours)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Food	93.06	86.13	93.56	88.30	93.64	89.25	94.34	90.18	94.72	91.39	95.56	91.80	95.75	92.89
Goods	95.31	94.85	95.77	95.30	95.59	95.12	95.88	95.34	95.93	95.87	96.05	96.10	96.55	96.44
Buildings	94.32	94.64	94.78	94.98	94.57	94.71	95.04	94.90	96.26	95.80	96.53	96.04	96.86	96.85
Animal	93.56	92.85	94.05	93.32	94.29	93.55	94.76	93.87	95.03	94.06	95.62	94.78	95.92	95.62
Human	91.55	91.34	92.04	92.13	91.75	91.87	92.27	92.16	92.54	92.31	92.67	92.53	92.74	92.74
Plant	93.05	93.40	93.85	93.71	93.50	93.37	94.38	94.21	95.04	94.97	95.67	95.30	97.02	96.97
Scenery	91.87	91.03	92.33	91.87	92.15	91.39	92.76	91.98	93.17	92.38	93.63	92.94	94.57	94.15
Average	93.25	92.03	93.77	92.80	93.64	92.75	94.02	93.23	94.67	93.83	95.10	94.21	**95.63**	**95.09**

Table 6
The experimental results of subtask aspect extraction over seven domains.

	Text		Image		Multimodal			
	BiLSTM		ResNet50		Concatenation		Interaction	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Food	91.02	89.95	45.14	3.74	89.24	88.51	90.72	91.11
Goods	83.24	81.27	59.02	7.63	85.24	82.57	85.68	82.79
Buildings	91.96	92.13	47.74	4.14	91.52	91.75	91.95	92.00
Animal	86.70	85.05	56.39	5.06	88.28	86.17	91.81	89.46
Human	77.04	68.39	41.19	4.92	77.38	71.27	78.52	69.49
Plant	85.36	85.75	55.36	5.49	88.43	88.49	96.07	95.49
Scenery	78.47	78.44	57.88	10.89	79.75	79.51	81.66	80.80
Average	84.83	83.00	51.82	5.98	85.69	84.04	88.06	85.88

Table 7
The experimental results of aspect polarity prediction over the whole dataset. "(w/o) Adv" means the corresponding model without adversarial training.

		Acc.	F1
Image	Concatenation	78.36	77.01
	Interaction	78.80	77.23
Text	Concatenation	88.47	87.46
	Interaction	89.48	88.78
MMAP	Concatenation	89.85	89.28
	Interaction	**90.64**	**90.14**
MMAP (w/o Adv)	Concatenation	88.24	87.95
	Interaction	89.43	89.73

ity of our MASAD dataset. Moreover, the performance over the seven domains differs significantly. Thus, integrating extra knowledge or transfer knowledge from other domains is an interesting point to investigate.

6.5. Discussions

First, we do some ablation studies to verify the effectiveness of each parts of our models.

- **Effectiveness of Interaction.** 1) the inter-interaction between the aspect and the multimodal information is important for this task. To be specific, we can find that the performance of the models with interaction is better than the ones with concatenation (Table 4). For example, the aspect "dog" can help the model to find the object "dog" in the image and find the word "dog" in the text (Fig. 1). The expression of the aspect and the words near the aspect (e.g., best, love) are important for sentiment prediction. 2) the intra-interaction between the text and image also plays a key role in this task. It is observed that our models with Interaction_{ITA} perform better than $\text{Concatenation}^{ITA}$ (Table 8). For instance, the word "running" in the text expresses the dog's action in the image.
- **Effectiveness of Adversarial Training.** From Table 4 and Table 7, we observe that the models with our adversarial training strategy performs better than the corresponding one without adversarial training. All these observations show that our adversarial training strategy can align the feature representation of the text and image effectively.

Second, we investigate the complexity and convergence of our proposed MMAP model (Table 9).

- **Convergence Analysis.** From Table 9, we find that our proposed MMAP model converges faster than the MIMN model. In addition, our model obtains the better performance than MIMN with the same epoch number. Also, these indicate that our model has a better convergence rate than the MIMN model.
- **Complexity Analysis.** We also calculate the parameters of these two models. The parameters of these two models are similar, which denotes that our MMAP model has a similar space complexity with the MIMN model. Since our MMAP model converges faster than MIMN, the time complexity of MMAP is smaller than it.

7. Conclusion and future work

In this paper, we focus on the new task named multimodal aspect-based sentiment analysis (MASA), which performs aspect-based sentiment analysis based on the multimodal content on social media platforms. We propose a multimodal interaction model for this task, which learns the interaction between the aspect, text, and image effectively. Then we present and release a large-scale multimodal aspect-based sentiment analysis dataset named MASAD. The MASAD involves 57 aspects with seven domains and contains 38 k samples with image-text pair. This dataset provides a new point of view for aspect-based sentiment classification, and also a new benchmark for MASA. We believe this

Table 8
The experimental results of subtask aspect polarity prediction over seven domains with text only, image only and combination of them. The mask $_{TA}$ and $_{IA}$ denote the concatenation/interaction for text and aspect, image and aspect respectively, $_{ITA}$ indicates the concatenation/interaction for image, text and aspect.

	Text				Image				Multimodal			
	$\text{Concatenation}_{TA}$		Interaction_{TA}		$\text{Concatenation}_{IA}$		Interaction_{IA}		$\text{Concatenation}_{ITA}$		Interaction_{ITA}	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Food	92.50	85.65	95.31	91.93	87.14	78.60	87.29	78.29	95.60	92.69	95.75	92.89
Goods	94.88	94.74	96.64	96.53	88.07	87.61	88.78	88.45	95.06	94.88	96.55	96.44
Buildings	93.88	93.81	96.01	96.00	83.53	83.53	85.85	85.83	96.01	95.99	96.86	96.85
Animal	93.43	92.90	93.55	92.90	79.42	77.06	79.59	77.52	94.25	93.75	95.92	95.62
Human	92.40	92.39	93.54	93.53	85.66	85.65	85.97	85.95	91.72	91.70	92.74	92.74
Plant	92.85	92.69	92.93	92.84	83.16	82.85	84.38	84.23	94.83	94.73	97.02	96.97
Scenery	91.82	90.87	93.96	93.43	82.95	81.25	82.95	81.46	94.11	93.64	94.57	94.15
Average	93.11	91.86	94.56	93.88	84.28	82.36	84.97	83.10	94.51	93.91	95.63	95.09

Table 9
The complexity and convergence of our MMAP model and MIMN. We report the accuracy of test set over first five epoches and the parameters of the models.

	Complexity	Convergence (Acc.)				
	Parameters	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
MIMN	26 M	84.51	85.71	86.34	86.18	87.62
MMAP (Ours)	25 M	86.53	87.09	88.73	89.33	89.87

dataset will push the state-of-the-art in MASA. We also propose several strong baselines for this task. Extensive experiments have been conducted on the MASAD dataset, showing high generalization capability of models trained on the proposed dataset and the benefit of using multiple visual modalities.

We can conclude from the experimental results and analysis that MASA is more challenging than textual aspect-based sentiment analysis. We believe the following research directions are worth studying: (1) Designing more expressive model architectures for learning the inter-multimodal information; (2) Exploring how to extract the aspect and predict the sentiment jointly; (3) Leveraging extra knowledge or transferring the knowledge from similar domains to improve the performance of MASA; (4) Expanding the dataset with multiple linguistics is an under-explored area where more work is expected.

CRediT authorship contribution statement

Jie Zhou: Writing - original draft, Conceptualization, Methodology, Software. **Jiabao Zhao:** Data curation, Methodology, Software. **Jimmy Xiangji Huang:** Visualization, Investigation, Supervision. **Qinmin Vivian Hu:** Methodology, Investigation. **Liang He:** Methodology, Validation, Supervision.

Declaration of Competing Interest

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

Acknowledgements

We greatly appreciate anonymous reviewers and the associate editor for their valuable and high quality comments that greatly helped to improve the quality of this article. This research is funded by the Science and Technology Commission of Shanghai Municipality (19511120200). This research is also supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the York Research Chairs (YRC) program. Jie Zhou and Jiabao Zhao are first co-authors with equal contribution. The corresponding authors is Liang He.

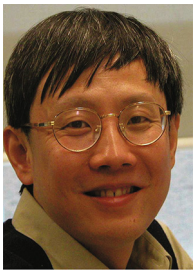
References

[1] Y. Lu, C. Zhai, Opinion integration through semi-supervised topic modeling, Proceedings of WWW (2008) 121–130.
[2] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, S.-F. Chang, Visual affect around the world: a large-scale multilingual visual sentiment ontology, in: Proceedings of ACMMM ACM, 2015, pp. 159–168.
[3] B. Liu, Sentiment analysis and opinion mining, Synthesis Lect Human Lang Technol 5 (2012) 1–167.
[4] J. Zhou, J.X. Huang, Q. Chen, Q.V. Hu, T. Wang, L. He, Deep learning for aspect-level sentiment classification: Survey, vision and challenges, IEEE Access (2019) 78454–78483.
[5] S. Manandhar, Semeval-2014 task 4: Aspect based sentiment analysis, Proceedings of SemEval (2014).
[6] M. Mitchell, J. Aguilar, T. Wilson, B. Van Durme, Open domain targeted sentiment, Proceedings of EMNLP (2013) 1643–1654.
[7] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, Semeval-2015 task 12: Aspect based sentiment analysis, Proceedings of SemEval (2015) 486–495.
[8] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, et al., Semeval-2016 task 5: Aspect based sentiment analysis, Proceedings of SemEval (2016) 19–30.
[9] M. Saeidi, G. Bouchard, M. Liakata, S. Riedel, Sentihood, Targeted aspect based sentiment analysis dataset for urban neighbourhoods, Proceedings of COLING (2016) 1546–1556.
[10] N. Xu, W. Mao, G. Chen, Multi-interactive memory network for aspect based multimodal sentiment analysis (2019) 371–378.
[11] J. Yu, J. Jiang, Adapting BERT for target-oriented multimodal sentiment classification (2019) 5408–5414.
[12] J. Zhou, J.X. Huang, Q.V. Hu, L. He, Modeling multi-aspect relationship with joint learning for aspect-level sentiment classification, in: International Conference on Database Systems for Advanced Applications, Springer, 2020, pp. 786–802.
[13] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, Proceedings of ACL (2017) 873–883.
[14] J. Zhou, J. Tian, R. Wang, Y. Wu, W. Xiao, L. He, Sentix, A sentiment-aware pre-trained model for cross-domain sentiment analysis, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 568–579.
[15] S. Kiritchenko, X. Zhu, C. Cherry, S. Mohammad, Nrc-canada-2014 Detecting aspects and sentiment in customer reviews, Proceedings of SemEval (2014) 437–442.
[16] D.-T. Vo, Y. Zhang, Target-dependent twitter sentiment classification with rich automatic features, Proceedings of IJCAI (2015) 1347–1353.
[17] D. Tang, B. Qin, X. Feng, T. Liu, Effective lstms for target-dependent sentiment classification, Proceedings of COLING (2016) 3298–3307.
[18] R. He, W.S. Lee, H.T. Ng, D. Dahlmeier, Exploiting document knowledge for aspect-level sentiment classification, Proceedings of ACL (2018) 579–585.

[19] X. Li, W. Lam, Deep multi-task learning for aspect term extraction with memory interaction, Proceedings of EMNLP (2017) 2886–2892.
[20] C. Fan, Q. Gao, J. Du, L. Gui, R. Xu, K.-F. Wong, Convolution-based memory network for aspect-based sentiment analysis, in: Proceedings of SIGIR ACM, 2018, pp. 1161–1164.
[21] J. Liu, Y. Zhang, Attention modeling for targeted sentiment, in: Proceedings of EACL, volume 2, 2017, pp. 572–577..
[22] S. Wang, S. Mazumder, B. Liu, M. Zhou, Y. Chang, Target-sensitive memory networks for aspect sentiment classification, Proceedings of ACL 1 (2018) 957–967.
[23] W. Xue, T. Li, Aspect based sentiment analysis with gated convolutional networks, in: Proceedings of ACL, Association for Computational Linguistics, 2018, pp. 2514–2523..
[24] J. Zhou, Q. Chen, J.X. Huang, Q.V. Hu, L. He, Position-aware hierarchical transfer model for aspect-level sentiment classification, Inf. Sci. 513 (2020) 1–16.
[25] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014)..
[26] Y. Wang, M. Huang, L. Zhao, et al., Attention-based lstm for aspect-level sentiment classification, Proceedings of EMNLP (2016) 606–615.
[27] D. Ma, S. Li, X. Zhang, H. Wang, Interactive attention networks for aspect-level sentiment classification, Proceedings of IJCAI (2017) 4068–4074.
[28] P. Chen, Z. Sun, L. Bing, W. Yang, Recurrent attention network on memory for aspect sentiment analysis, Proceedings of EMNLP (2017) 452–461.
[29] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, Z. Wu, Content attention model for aspect based sentiment analysis, in: Proceedings of WWW, International World Wide Web Conferences Steering Committee, 2018, pp. 1023–1032..
[30] J. Wang, J. Li, S. Li, Y. Kang, M. Zhang, L. Si, G. Zhou, Aspect sentiment classification with both word-level and clause-level attention networks, Proceedings of IJCAI (2018) 4439–4445.
[31] F. Zhao, Z. Wu, X. Dai, Attention transfer network for aspect-level sentiment classification, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 811–821.
[32] J. Zhou, J.X. Huang, Q.V. Hu, L. He, Sk-gcn: Modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification, Knowl.-Based Syst. 205 (2020) 106292.
[33] Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm, Proceedings of AAAI (2018) 5876–5883.
[34] H. Xu, B. Liu, L. Shu, S.Y. Philip, Bert post-training for review reading comprehension and aspect-based sentiment analysis, 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), 2019, pp. 2324–2335.
[35] M.H. Phan, P.O. Ogunbona, Modelling context and syntactical features for aspect-based sentiment analysis, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3211–3220.
[36] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, K. Xu, Adaptive recursive neural network for target-dependent twitter sentiment classification, Proceedings of ACL 2 (2014) 49–54.
[37] Z. Fan, Z. Wu, X. Dai, S. Huang, J. Chen, Target-oriented opinion words extraction with target-fused neural sequence labeling, 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), 2019, pp. 2509–2518.
[38] Q. Jiang, L. Chen, R. Xu, X. Ao, M. Yang, A challenge dataset and effective models for aspect-based sentiment analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6281–6286.
[39] M. Chen, S. Wang, P.P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of ICMI ACM, 2017, pp. 163–171.
[40] S. Poria, H. Peng, A. Hussain, N. Howard, E. Cambria, Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis, Neurocomputing 261 (2017) 217–230.
[41] R. Kaur, S. Kautish, Multimodal sentiment analysis: A survey and comparison, IJSSMET 10 (2019) 38–58.
[42] D. Ghosal, M.S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, P. Bhattacharyya, Contextual inter-modal attention for multi-modal sentiment analysis, Proceedings of EMNLP (2018) 3454–3466.
[43] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, L.-P. Morency, Multi-attention recurrent network for human communication comprehension, Proceedings of AAAI (2018).
[44] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, , Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: Proceedings of ACMMM ACM, 2013, pp. 223–232.
[45] X. Jin, A. Gallagher, L. Cao, J. Luo, J. Han, The wisdom of social multimedia: using flickr for prediction and forecast, in: Proceedings of ACMMM, ACM, 2010, pp. 1235–1244.
[46] K. Krippendorff, Computing krippendorff's alpha-reliability (2011)..
[47] J. Pennington, R. Socher, C. Manning, Glove, Global vectors for word representation, Proceedings of EMNLP (2014) 1532–1543.
[48] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional lstm and other neural network architectures, Neural Networks 18 (2005) 602–610.
[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of CVPR (2016) 770–778.
[50] V. Gajarla, A. Gupta, Emotion detection and sentiment analysis of images, Georgia Institute of Technology, 2015.
[51] X. Chen, Z. Shi, X. Qiu, X.-J. Huang, Adversarial multi-criteria learning for chinese word segmentation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1193–1203.
[52] D. Tang, B. Qin, T. Liu, Aspect level sentiment classification with deep memory network, Proceedings of EMNLP (2016) 214–224.
[53] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of ICLR, volume 5, 2015..
[54] X. Li, L. Bing, W. Lam, B. Shi, Transformation networks for target-oriented sentiment classification, Proceedings of ACL (2018) 946–956.

Jie Zhou is currently working toward the PhD degree at the Department of Computer Science and Technology, East China Normal University, China. His research interest includes sentiment analysis, aspect-level sentiment classification, retrieval model, and neural networks. He was awarded a scholarship from the China Scholarship Council, and received Top-3 in the KDD CUP Competition several times. Since 2016, he has published more than 10 referred papers in international conferences or journals, such as AAAI, Information Sciences, DASFAA, ICME.

Jiabao Zhao is currently working toward the PhD degree at the Department of Computer Science and Technology, East China Normal University, China. Her research interest includes few-shot learning, meta-learning, and neural networks. Since 2017, she has published several referred papers in international conferences or journals, such as ICME.



Jimmy Xiangji Huang received the Ph.D degree in information science from the City, University of London and was then a post doctoral fellow in the School of Computer Science, University of Waterloo, Canada. He is now a York Research Chair professor and the director of Information Retrieval & Knowledge Management Research Lab (IRLab), York University. He joined York University as an assistant professor, in 2003. He was awarded tenure and promoted to full professor in 2006 and 2011 respectively. He received the Deans Award for Outstanding Research in 2006, an Early Researcher Award, formerly the Premiers Research Excellence Award in 2007, the Petro Canada Young Innovators Award in 2008, the SHARCNET Research Fellowship Award in 2009, the Best Paper Award at the 32nd European Conference on Information Retrieval (ECIR 2010), United Kingdom, and LA&PS Award for Distinction in Research, Creativity, and Scholarship (established researcher) in 2015. Since 2003, he has published more than 230 refereed papers in journals (such as the ACM Transactions on Information Systems, the Journal of American Society for Information Science and Technology, Information Processing & Management, the IEEE Transactions on Knowledge and Data Engineering, Information Sciences, Information Retrieval, BMC Bioinformatics, BMC Genomics and BMC Medical Genomics), book chapters, and international conference proceedings (such as ACM SIGIR, ACM CIKM, KDD, ACL, COLING, IEEE ICDM, IJCAI and AAAI). He is a senior member of the IEEE & ACM Distinguished Scientist.

Qinmin Vivian Hu received her Ph.D. degree in Computer Science from York University, Toronto, Canada. She is now an Assistant Professor in the Department of Computer Science, Ryerson University, Toronto, Canada. Before that, she was an Associate Professor at East China Normal University, Shanghai, China, and a Postdoctoral Fellow at the MRI research facility, the Wayne State University, USA. She won the National NSERC Postdoctoral Fellowship as one of best Ph.D fellows in Canada in 2013. She has published more than 30 referred papers in top tier journals (i.e. IEEE Transactions on Knowledge and Data Engineering, ACM Transactions on Intelligent Systems and Technology) and conferences (i.e. AAAI, ACM SIGIR).

Liang He received his bachelor's degree and PhD degree from the Department of Computer Science and Technology, East China Normal University, China. He is now a professor and the director of the Department of Computer Science and Technology, East China Normal University. His current research interest includes knowledge processing, user behavior analysis, and context-aware computing. He has been awarded the Star of the Talent in Shanghai. He is also a council member of the Shanghai Computer Society, a member of the Academic Committee, the director of the technical committee of Shanghai Engineering Research Center of Intelligent Service Robot, and a technology foresight expert of the Shanghai Science and Technology in focus areas. He has received the Shanghai Science and Technology Progress Award for 5 times, and won the first prize in 2013 and the second prize in 2015. He has obtained more than 10 patents, and published 2 monographs and more than 70 refereed papers in national and international journals and conference proceedings.