Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection

Xincheng Ju[1], **Dong Zhang**[1]*, **Rong Xiao**[2], **Junhui Li**[1],
Shoushan Li[1], **Min Zhang** [1], **Guodong Zhou**[1]
[1] School of Computer Science and Technology, Soochow University, China
[2] Alibaba Group, China
xcju@stu.suda.edu.cn, xiaorong.xr@taobao.com,
{dzhang, jhli, lishoushan, minzhang, gdzhou}@suda.edu.cn

Abstract

Aspect terms extraction (ATE) and aspect sentiment classification (ASC) are two fundamental and fine-grained sub-tasks in aspect-level sentiment analysis (ALSA). In the textual analysis, jointly extracting both aspect terms and sentiment polarities has been drawn much attention due to the better applications than individual sub-task. However, in the multi-modal scenario, the existing studies are limited to handle each sub-task independently, which fails to model the innate connection between the above two objectives and ignores the better applications. Therefore, in this paper, we are the first to jointly perform multi-modal ATE (MATE) and multi-modal ASC (MASC), and we propose a multi-modal joint learning approach with auxiliary cross-modal relation detection for multi-modal aspect-level sentiment analysis (MALSA). Specifically, we first build an auxiliary text-image relation detection module to control the proper exploitation of visual information. Second, we adopt the hierarchical framework to bridge the multi-modal connection between MATE and MASC, as well as separately visual guiding for each sub module. Finally, we can obtain all aspect-level sentiment polarities dependent on the jointly extracted specific aspects. Extensive experiments show the effectiveness of our approach against the joint textual approaches, pipeline and collapsed multi-modal approaches.

1 Introduction

Multi-modal aspect-level (aka target-oriented) sentiment analysis (MALSA) is an important and fine-grained task in multi-modal sentiment analysis (MSA). Previous studies normally cast MALSA in social media as two independent sub-tasks: Multi-modal Aspect Terms Extraction (MATE) and Multi-modal Aspect Sentiment Classification (MASC). First, MATE aims to detect a set of all potential aspect terms from a free text with its accompanying image (Wu et al., 2020a). Second, MASC aims to classify the sentiment polarity of a multi-modal post towards a given aspect in textual modality (Yu and Jiang, 2019).

[**NBA**]$_{\text{Neu}}$: [**Spurs**]$_{\text{Pos}}$ rout [**Thunder**]$_{\text{Neg}}$ | Tempo Sports

(a)

RT @ funnytwittingg : [**OBAMA**]$_{\text{Neg}}$ TO [**ISRAEL**]$_{\text{Neu}}$? [**OBAMA**]$_{\text{Neg}}$ TO [**UKRAINE**] ? [**OBAMA**]$_{\text{Neg}}$ TO [**USA**]$_{\text{Neu}}$?

(b)

Figure 1: Two examples for joint multi-modal aspect-sentiment analysis.

To better satisfy the practical applications, the aspect term-polarity co-extraction, which solves ATE and ASC simultaneously, receives much attention recently in a textual scenario (Wan et al., 2020; Chen and Qian, 2020b; Ying et al., 2020). However, to our best knowledge, in the multi-modal scenario, the joint MATE and MASC, i.e., joint multi-modal aspect-sentiment analysis (JMASA), have never been investigated so far. For this joint multi-modal task, we believe that there exist the following challenges at least.

On the one hand, visual modality may provide no clues for one of sub-tasks. For example, in Figure 1(a), since the image shows most of the content described in the text, and we can't infer from the image which team has an advantage at first glance. While, a direct understanding of the text (e.g., the word "*rout*") seems to be able to judge the sentiment of "*Spurs*" and "*Thunder*". Thus this image does not add to the text tweet meaning (Vempala and Preotiuc-Pietro, 2019). On the contrary, in Figure 1(b), the information of textual modality is

* Corresponding Author

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4395–4405
November 7–11, 2021.

quite limited so that we cannot directly infer the sentiment towards one aspect. While, the visual modality provides rich clues (e.g., differential expressions) to help us predict the correct sentiment of "*OBAMA*". Therefore, a well-behaved approach should determine whether the visual information adds to the textual modality (cross-modal relation detection) and how much visual information contributes to text.

On the other hand, the characteristics of the two multi-modal sub-tasks are different: one is sequence labeling problem, the other is aspect-dependent classification problem. Different tasks seem to focus on different image information. For example, in Figure 1(b), towards first sub-task MATE, if we can attend to some coarse-grained concepts (e.g., silhouette of human face, *Person* label) in the image, it is enough and effective to help identify the name "*OBAMA*" in the text as an aspect. Towards second sub-task MASC, we should attend to the details (e.g., different facial expressions) of some regions, so that we can judge the accurate sentiment dependent on a specific aspect "*OBAMA*". Therefore, a well-behaved approach should separately mine the visual information for these two sub-tasks instead of collapsed tagging with the same visual feeding.

To handle the above challenges, we propose a multi-modal joint learning approach with auxiliary cross-modal relation detection, namely JML. Specifically, we first design a module of auxiliary cross-modal relation detection to control whether the image adds to the text meaning. Second, we leverage the joint hierarchical framework to separately attend to the effective visual information for each sub-task instead of collapsed tagging framework. Finally, we can obtain all potential aspect term-polarity pairs. Extensive experiments and analysis on two multi-modal datasets in Twitter show that our approach performs significantly better than text-based joint approaches and collapsed multi-modal joint approaches.

2 Related Work

In the past five years, text-based aspect-level sentiment analysis has drawn much attention (Luo et al., 2019; Chen and Qian, 2019; Zhang and Qian, 2020; Zheng et al., 2020; Tulkens and van Cranenburgh, 2020; Akhtar et al., 2020). While, multi-modal target-oriented sentiment analysis has become more and more vital because of its urgent need to be applied to the industry recently (Akhtar et al., 2019; Zadeh et al., 2020; Sun et al., 2021a; Tang et al., 2019; Zhang et al., 2020b, 2021a). In the following, we mainly overview the limited studies of multi-modal aspect terms extraction and multi-modal aspect sentiment classification on text and image modalities. Besides, we also introduce some representative studies for text-based joint aspect terms extraction and sentiment polarity classification.

Multi-modal Aspect Terms Extraction (MATE). Sequence labeling approaches are typically employed for this sub-task(Ma et al., 2019; Chen and Qian, 2020a; Karamanolakis et al., 2019). But it is challenging to bridge the gap between text and image. Several related studies with focus on named entity recognition propose to leverage the whole image information by ResNet encoding to augment each word representation, such as (Moon et al., 2018; Zhang et al., 2018) upon RNN, (Yu et al., 2020b) upon Transformer and (Zhang et al., 2021b) on GNN. Besides, several related studies propose to leveraging the fine-grained visual information by object detection, such as (Wu et al., 2020a,b)

However, all the above studies completely ignore the sentiment polarity analysis dependent on the detected target, which has great facilitates in practical applications, such as e-commerce. Different from them, we propose to jointly perform the corresponding sentiment classification besides aspect terms extraction in a multi-modal scenario. Note that we propose a multi-modal joint learning approach to improve the performance of both MATE and MASC.

Multi-modal Aspect Sentiment Classification (MASC). Different from text-based aspect sentiment classification (Sundararaman et al., 2020; Ji et al., 2020; Liang et al., 2020b,a), it is challenging to effectively fuse the textual and visual information. As a pioneer, Xu et al. (2019) collect a benchmark Chinese dataset from a digital product review platform for multi-modal aspect-level sentiment analysis and propose a multi-interactive memory network to iteratively fuse the textual and visual representations.

Recently, Yu and Jiang (2019) annotate two datasets in Twitter for multi-modal target-oriented (aka aspect-level) sentiment classification and leverage BERT as backbone to effectively combine both textual and visual modalities. In the same period,

Yu et al. (2020a) propose a target-sensitive attention and fusion network to address both text-based and multi-modal target-oriented sentiment classification.

However, all the above studies assume that the aspect or target has been given, which is limited to some applications. Different from them, we propose to jointly perform aspect terms extraction besides the corresponding sentiment classification in a multi-modal scenario. Note that we also propose a multi-modal joint learning approach to improve the performance of both MATE and MASC.

Text-based Joint Aspect Terms Extraction and Sentiment Classification. Some studies (Zhang et al., 2020a) have attempted to solve both sub-tasks in a more integrated way, by jointly extracting aspect terms and predicting their sentiment polarities. The most recent and representative are a span-based extract-then-classify approach (Hu et al., 2019) and a directed GCN with syntactic information (Chen et al., 2020).

However, all the above studies can not model the visual guidance for both sub-tasks. Different from them, we propose a multi-modal joint framework to handle both MATE and MASC.

3 Joint Multi-modal Aspect-Sentiment Analysis

In this section, we introduce our approach for multi-modal aspect terms extraction and aspect sentiment classification jointly. In the following, we first formalize this joint task, then introduce the module of text-image relation detection, finally give the details of our hierarchical framework for multi-modal learning.

Task Definition We define the following notations, used throughout the paper. Let $\mathcal{D} = \{(X_n, I_n, A_n, S_n)\}_{n=1}^{N}$ be the set of data samples. Given a word sequence $X = \{x_1, x_2, \cdots, x_k\}$ with length k and an image I, the joint task is to extract a aspect terms list $A = \{a_1, a_2, \cdots, a_m\}$ and classify the aspect sentiment list $S = \{s_1, s_2, \cdots, s_m\}$ simultaneously, where m denotes the number of aspects. Note that the word embeddings are obtained by pre-processing via BERT (Devlin et al., 2019) due to its excellent ability of textual representation, meanwhile the image region embeddings are obtained by pre-processing via ResNet (He et al., 2016) due to its excellent ability of visual representation.

3.1 Cross-modal Relation Detection

Unlike traditional approaches, which take visual information into consideration completely and ignore whether image can bring benefits to text, we incorporate the image-text relation into the model and only retain the auxiliary visual information towards the text. Therefore, we build a relation module by pre-training to properly exploit visual modality for our joint multi-modal tasks. The cross-modal relation detection module is shown in bottom right corner of Figure 2.

We employ TRC dataset (Vempala and Preotiuc-Pietro, 2019) for text-image relation detection to control whether image adds to the text meaning. Table 1 shows the types of text-image relations and statics of the TRC dataset.

Module Design. we first involve two raw modalities into pre-trained module of BERT and ResNet respectively, noting that the pre-trained module involved in cross-modal relation detection module independently. Then, we incorporate two modal representation into a self-attention block to capture intra-modal interactions for each modality. After that, we put output states into the cross-attention block capture inter-modal interactions for text and image. Formally,

$$H_o = \mathrm{ATT}_{\mathrm{self}}(O_{rel}) \quad (1)$$

$$H_x = \mathrm{ATT}_{\mathrm{self}}(T_{rel}) \quad (2)$$

$$H_{o\to x} = \mathrm{ATT}_{\mathrm{cross}}(H_o, H_x) \quad (3)$$

$$H_{x\to o} = \mathrm{ATT}_{\mathrm{cross}}(H_x, H_o) \quad (4)$$

where $\mathrm{ATT}_{\mathrm{self}}$ denotes self-modal multi-head attention as (Vaswani et al., 2017), and $\mathrm{ATT}_{\mathrm{cross}}$ denotes the cross-modal multi-head attention as (Ju et al., 2020). O_{rel} and T_{rel} are pre-trained embedding of image I and text X.

Finally, we obtain the relation probabilities through a feed-forward neural network and a softmax activation function as follows:

$$p_r = \mathrm{softmax}(W_2 \tanh(W_1 H)) \quad (5)$$

where $W_1 \in \mathbb{R}^{4*d_m \times d_m}$ and $W_2 \in \mathbb{R}^{d_m \times 2}$ are two trainable parameter matrices. H means the concatenation of H_o, H_x, $H_{o\to x}$ and $H_{x\to o}$. Since the relation score can also be binary: 0 or 1, we calculated by equation similarly to equation 5,but score $p_r < 0.5 = 0$, $p < 0.5$. Then we try both soft and hard relations to guide our multi-modal joint tasks.

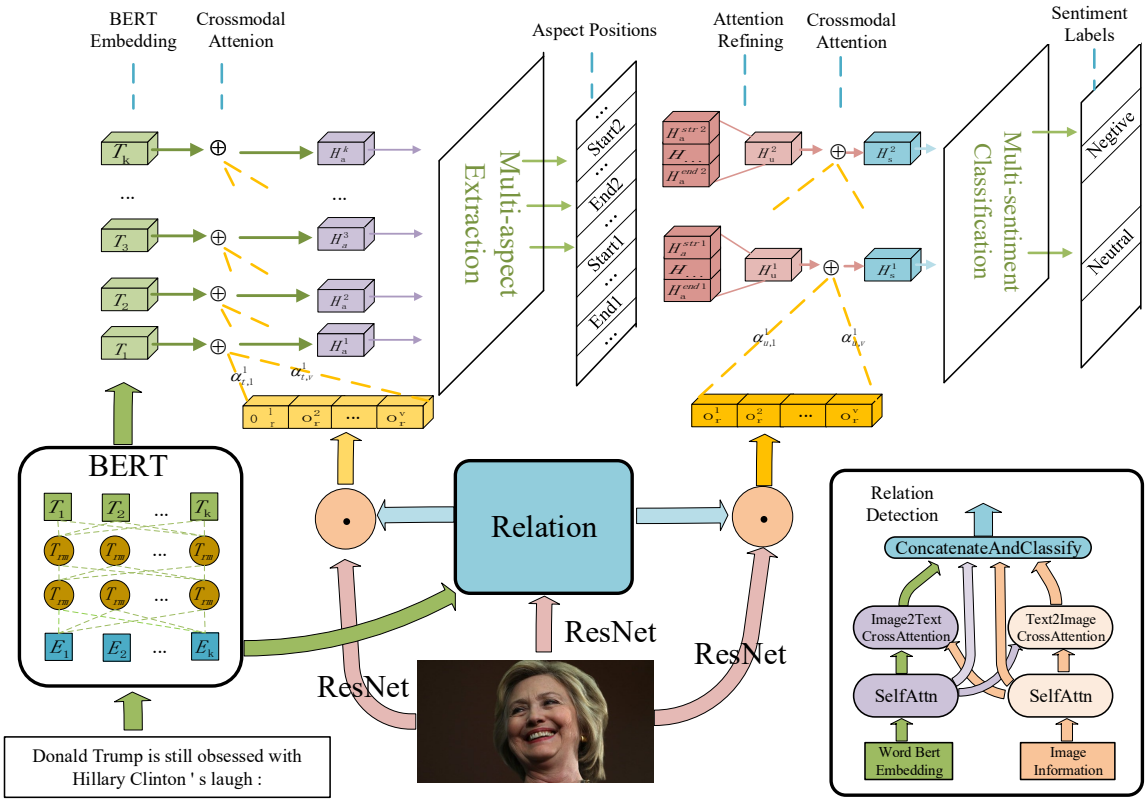

Figure 2: The overview of our proposed JML.

Relation Loss. Let $\mathcal{D}_r = \{r\}_{n=1}^{N} = \{< text^{(i)}, image^{(i)} >\}_{i=1}^{N}$ be a set of text-image pairs for TRC training. The loss $\mathcal{L}_r$ of binary relation classification is calculated by cross entropy:

$$\mathcal{L}_r = -\sum_{i=1}^{N} \log(p_r(r^{(i)})) \tag{6}$$

where $p_r(x)$ is the probability for correct classification and the probability is calculated by softmax.

3.2 Multi-modal Aspect Terms Extraction

The left part of Figure 2 shows the architecture of multi-modal aspect terms extraction. We first leverage the text-image relation to control the visual input, then make the textual and visual information perform mutual attention.

$$G_r = \text{RelDet}(X, I) \tag{7}$$
$$O_r = G_r \cdot O \tag{8}$$

where RelDet(,) denotes the relation detection module with inputs X and I. O is the output of another ResNet for I, applied for our main task. G_r is the relation score. In this stage, we use the mask gate G_r to control the additive visual clues.

Subsequently, we make the text attend to the effective visual information of the first sub-task MATE. Defined as follows:

$$C_a = T \oplus \text{ATT}_{\text{cross}}(T, O_r) \tag{9}$$
$$H_a = W_a C_a + b_a \tag{10}$$

where $\oplus$ denotes the element-wise addition and $W_a \in \mathbb{R}^{d_m * d_m}$. T is the output of another BERT for X, applied for our main task.

Instead of finding aspects via BIO sequence tagging approaches, we identify candidate aspects by its start and end positions in the sentence, inspired by previous research (Wang and Jiang, 2017; Hu et al., 2019), due to the huge search space and the inconsistence of multi-word sentiment. From the above step, we obtain the unnormalized score as well as the probability distribution of the start position as:

$$g^{str} = W_s H_a + b_s \tag{11}$$
$$p^{str} = \text{softmax}(g^{str}) \tag{12}$$

where $W_s \in \mathbb{R}^{d_m}$ is a trainable weight vector. Correspondingly, we can obtain the end position probability along with its confidence score by:

$$g^{end} = W_e H_a + b_e \tag{13}$$
$$p^{end} = \text{softmax}(g^{end}) \tag{14}$$

During training, considerate that each sentence may contain multiple aspects, we label the span boundaries for all aspect entities in the A. After that, we obtain a vector $y^s \in \mathbb{R}^k$, where each element y_i^s indicates whether the i-th position is the start of an aspect, and also we get another vector y^e for labeling the end positions.

3.3 Multi-modal Aspect Sentiment Classification

Traditionally, aspect sentiment classification with aspect focus on using either sequence tagging methods or sophisticated neural networks that separately encode the target and the sentence. Instead, we propose to obtain the summarized representation from the upper layer cross-modal state H_a based on position vectors (y^s and y^e). Then, a feed-forward neural network is used to predict the sentiment polarity as shown in Figure 2 (upper right corner).

Inspired by the upper network, we receive a multiple aspect span list from y^s and y^e. Specially, given an aspect span a, we summarize hidden state representation H_a in its corresponding bound (s_i, e_i) as a vector H_u^i with the attention mechanism (Bahdanau et al., 2015). Formally:

$$m = W_m H_a^{[s_i:e_i]} + b_m \quad (15)$$

$$\alpha_t = \frac{\exp(m_t)}{\sum_{t \in m} \exp(m_t)} \quad (16)$$

$$H_u^i = \sum_{t \in \alpha} \alpha_t H_a^{s_i+t} \quad (17)$$

where $W_m \in \mathbb{R}^{d_m}$ is a trainable weight vector.

In addition, we integrate visual representation O_r in formula (8) into span vector set H_u with assistance of relation gate G_r. Similar to formula (9-10), cross-modal multi-head attention mechanism is used to modal fusion:

$$C_s = H_u \oplus \mathrm{ATT}_{\mathrm{cross}}(H_u, O_r) \quad (18)$$

$$H_s = W_u C_s + b_u \quad (19)$$

where $W_u \in \mathbb{R}^{d_m \times d_m}$, and then we get $H_s \in \mathbb{R}^{m \times d_m}$ as final sentiment state set.

Furthermore, we obtain the polarity score by applying two linear transformations with a Tanh activation in between, and is normalized with a softmax function to output the polarity probability as:

$$g^p = W_p \tanh(W_v H_s) \quad (20)$$

$$p^p = \mathrm{softmax}(g^p) \quad (21)$$

where $W_p \in \mathbb{R}^{d_m \times \epsilon}$ and $W_v \in \mathbb{R}^{d_m \times d_m}$ are two trainable weight parameter matrices. ϵ in the number of sentiment classes.

3.4 Joint Loss

Since it is a joint task with aspect terms extraction and aspect sentiment classification, we calculate two different sets of loss simultaneously as follows:

$$\mathcal{L} = -\sum_{i=1}^{k} y_i^s \log(p_i^{str}) - \sum_{i=1}^{k} y_i^e \log(p_i^{end}) - \sum_{t=1}^{m} \sum_{i=1}^{\epsilon} y_{ti}^p \log(p_{ti}^p) \quad (22)$$

where y^s, y^e, y^p are one-hot labels indicating golden start, end positions, true sentiment polarity separately, and a, m are the number of sentence tokens, aspects respectively.

At inference time, we select the most suitable span(k,l)(k<l) with assist of position polarity (g^{str}, g^{end}) as final aspect prediction based on previous research (Hu et al., 2019). After that, the sentiment polarity probability is calculated for each candidate span and select the sentiment class with the maximum value in p^p.

4 Experimentation

In this section, we systematically evaluate our approach to aspect terms extraction and aspect sentiment classification.

4.1 Experimental Settings

Datasets. In the experiments, we use three datasets to evaluate the performance. One is the TRC dataset, and the other two are public Twitter datasets (i.e., Twitter2015 and Twitter2017) for MALSA. The detailed descriptions are as follows:

TRC dataset of Bloomberg LP (Vempala and Preotiuc-Pietro, 2019) In this tweets dataset, we select two types of text-image relation annotated by the authors, as shown in Table 1. "Image adds to the tweet meaning" focus on the usefulness of image

	R_1	R_2
Image adds to the tweet meaning	✓	✗
Percentage(%)	44.2	55.8

Table 1: Two relation types in the TRC dataset, representing that image adds to the tweet meaning or not respectively.

Dataset	Twitter-2015							Twitter-2017						
	Pos	Neu	Neg	Total	AS	Words	AL	Pos	Neu	Neg	Total	AS	Words	AL
Train	928	1883	368	3179	1.348	9023	16.72	1508	1638	416	3562	1.410	6027	16.21
Valid	303	670	149	1122	1.336	4238	16.74	515	517	144	1176	1.439	2922	16.37
Test	317	607	113	1037	1.354	3919	17.05	493	573	168	1234	1.450	3013	16.38

Table 2: The statistics summary of Twitter-2015 and Twitter-2017 datasets. Pos: Positive, Neu: Neutral, Neg: Negative, AL: Avg. Length, AS: Avg. Aspects

to the semantics of the tweet, especially suitable for our task. We follow the same split of 8:2 for train/test sets as in (Vempala and Preotiuc-Pietro, 2019).

Twitter dataset. As shown in Table 2, the dataset (i.e., Twitter2015 and Twitter2017) are provided by (Zhang et al., 2018) for multi-modal named entity recognition originally and annotated the sentiment polarity for each aspect by(Lu et al., 2018). we use this dataset for our joint task.

Implementation Details. We implement our approach via Pytorch toolkit (torch-1.1.0) with a piece of GTX 1080 Ti. The hidden size d_m in our model is 768 same to dim in BERT (Devlin et al., 2019). The number of heads in $\mathrm{ATT}_{\mathrm{self}}$ and $\mathrm{ATT}_{\mathrm{cross}}$ is 8.

During training, we train each model for a fixed number of epochs 50 and monitor its performance on the validation set. Once the training is finished,we select the model with the best F_1 score on the validation set as our final model and evaluate its performance on the test set. We adopt cross-entropy as the loss function and use the Adam (Kingma and Ba, 2015) optimization method to minimize the loss over the training data. To motivate future research, the code will be released via github[1]

Evaluation Metrics and Significance Test. In our study, we employ three evaluation metrics to measure the performances of different approaches to multi-modal aspect terms extraction and aspect sentiment classification jointly, i.e. micro F_1 measure (F_1), Precision(P) and Recall(R). Besides, through scipy[2],the paired t-test is performed to test the significance of the difference between two approaches, with a default significant level of 0.05. These metrics have been popularly used in some aspect extraction and sentiment classification problems.

4.2 Baselines

For a thorough comparison, we mainly compare four groups of baseline systems with our approach.

The first group are the most related approaches to multi-modal aspect terms extraction. 1) **RAN** (Wu et al., 2020a); a co-attention approach for aspect terms extraction in a multi-modal scenario. 2) **UMT** (Yu et al., 2020b); 3) **OSCGA** (Wu et al., 2020b), an NER approach in a multi-modal scenario based on object features with BIO tagging. Note that **UMT** and **OSCGA** focus on named entity recognition (NER) with BIO tagging in a multi-modal scenario, leveraging the representation ability of transformer and object-level fine-grained visual features, respectively.

The second group are the representative approaches to multi-modal aspect-dependent sentiment classification. 1) **TomBERT** (Yu and Jiang, 2019). 2) **ESAFN** (Yu et al., 2020a). Note that **TomBERT** is based on BERT, **ESAFN** is based on LSTM but explicitly models textual contexts.

The third group are text-based approaches of joint aspect terms extraction and aspect sentiment classification. 1) **SPAN** (Hu et al., 2019). 2) **D-GCN** (Chen et al., 2020). Note that **SPAN** also adopts a hierarchical framework but limited to textual scenario. **D-GCN** leverage syntactic information with GCN.

The fourth group are mainly multi-modal approaches for both sub-tasks. Since there exists no multi-modal approach for JMASA, we implement two pipeline approaches upon two representative studies of MATE and MASC and three collapsed tagging approaches. 1) **UMT+TomBERT**. 2) **OSCGA+TomBERT**. 3) **UMT-collapsed** (Yu et al., 2020b). 4) **OSCGA-collapsed** (Wu et al., 2020b). 5) **RpBERT** (Sun et al., 2021b). Note that **RpBERT** is a multi-modal multi-task approach for NER and text-image relation. Although it also

[1]https://github.com/MANLP-suda/JML.git.
[2]https://www.scipy.org/

Modality	Approaches	Twitter-2015			Twitter-2017		
		F	P	R	F	P	R
Text-based	SPAN (Hu et al., 2019)	53.8	53.7	53.9	60.6	59.6	61.7
	D-GCN (Chen et al., 2020)	59.4	58.3	58.8	64.1	64.2	64.1
Multi-modal Joint Task	UMT+TomBERT	59.8	58.4	61.3	62.4	62.3	62.4
	OSCGA+TomBERT	62.5	61.7	63.4	63.7	63.4	64.0
	UMT-collapse (Yu et al., 2020b)	61.0	60.4	61.6	60.8	60.0	61.7
	OSCGA-collapse (Wu et al., 2020b)	63.2	63.1	**63.7**	63.5	63.5	63.5
	RpBERT (Sun et al., 2021b)	48.0	49.3	46.9	56.2	57.0	55.4
	JML	**64.1**†‡	**65.0**	63.2	**66.0**†‡	**66.5**	**65.5**
	JML (hard)	62.8	63.9	61.8	63.7	64.1	63.2

Table 3: The performance comparison of different approaches, which simultaneously perform aspect terms extraction and aspect sentiment classification in both text-based and multi-modal scenarios. JML corresponds to soft relation, compared to JML(hard) with hard relation. The marker † refers to significant test $p\text{-}value < 0.05$ when comparing with **D-GCN**. The marker ‡ refers to significant test $p\text{-}value < 0.05$ when comparing with **OSCGA+TomBERT**. F: Micro F1, P: Precision, R:Recall

Approaches	F_1 score
(Lu et al., 2018)	81.0
RpBERT(Sun et al., 2021b)	88.1
JML	**89.8**

Table 4: Results of the text-image relation classification in F_1 score (%)

Approaches	Twitter-2015			Twitter-2017		
	F1	P	R	F1	P	R
RAN	81.0	80.5	81.5	90.3	90.7	90.0
UMT	79.7	77.8	81.7	86.7	86.7	86.8
OSCGA	81.9	81.7	**82.1**	90.4	90.2	90.7
JML-MATE	**82.4**	**83.6**	81.2	**91.4**	**92.0**	**90.7**

Table 5: Performance of multi-modal aspect terms extraction, compared with sub-task in our joint approach

leverages cross-modal relation, but it relies on collapsed tagging, which can not attend to different features for different multi-modal sub-tasks.

4.3 Experimental Results

Result of TRC. Table 4 shows the performance of our relation detection module on the test set of TRC data. The result shows that our attention-based visual-linguistic model equipped with BERT and ResNet outperforms that of (Lu et al., 2018) and **RpBERT**. F_1 score of our model on the test set of TRC data increases by 8.8% compared to (Lu et al., 2018) and 1.7% compared to RpBERT significantly, which demonstrated the effectiveness of this task.

For JMASA. Table 3 shows the results of different approaches in multi-modal scenarios, which simultaneously process the aspect terms extraction and aspect sentiment classification. From this table, we can observe that 1) text-based joint approaches perform much worse than multi-modal joint task approaches, suggesting that visual modality enriches representation to help correct predictions, rather than limited textual modality. 2) **UMT-collapse**, **RpBERT** and **OSCGA-collapse** perform much worse than our joint approach, owing to collapsed tagging with the same visual feeding, instead of separately mine the visual information for two subtasks. 3) **RpBERT** performs the worst in all baselines, which simultaneously process multiple tasks for text-image relation classification and visual-linguistic learning for aspect terms extraction and aspect sentiment classification, suggesting that a vanilla Bert-based model can not handle multiple tasks in the same time and greatly reduce task performance. 4) **JML(hard)** with a hard relation perform worse than its soft counterpart, indicating the wisdom of using soft image-text relation. 5) Among all the approaches, our proposed **JML** performs best in terms of almost all metrics. For instance, in terms of metric on Twitter-2017, our approach outperforms D-GCN by 1.9%, 2.3% and 1.4% with respect to $MicroF1$, $Precision$ and $Recall$, respectively. This is mainly because our approach with the joint framework leverages the indeed beneficial clues to two sub-task specially by cross-modal relation detection and cross-modal attention integration.

For MATE. Table 5 show the performance of

Approaches	Twitter-2015 Acc	Twitter-2017 Acc
TomBERT	74.0	70.9
ESAFN	70.9	65.5
JML-MASC	**78.7**	**72.7**

Table 6: Performance of multi-modal aspect-level sentiment classification, compared with sub-task in our joint approach

Approaches	Twitter-2015			Twitter-2017		
	F1	P	R	F1	P	R
JML (Full)	**64.1**	**65.0**	63.2	**66.0**	**66.5**	**65.5**
w/o Relation All	62.7	62.1	63.3	64.8	64.2	65.5
w/o Relation MAE	63.7	63.0	**64.3**	64.9	65.2	64.7
w/o Relation MASC	63.3	64.1	62.5	65.3	65.8	64.8
w/o Vision MAE	62.4	64.1	60.9	64.7	65.5	63.9
w/o Vision MASC	62.3	62.3	62.4	64.3	64.8	63.9

Table 7: The performance comparison of our full model JML and its ablated approaches.

different approaches, which only participate in multi-modal aspect terms extraction, compared with the sub-task performance in our joint approach. From this table, we can observe that 1) **UMT** performs the worst among all baselines, this is due to the fact that **SPAN** aligns text with object regions that show in an image and **OSCGA** combines object-level image information and character-level text information to predict aspects. 2) The sub-task performance in our joint approach performs better in most terms of metric, suggesting that our approach of joint framework promotes aspect terms extraction with the assistance of aspect sentiment information and relation-based visual modality.

For MASC. Table 6 shows the performance of different approaches, which only participate in multi-modal aspect sentiment classification, compared with the sub-task performance in our joint approach. From this table, we can observe that 1) **TomBERT** performs better than **ESAFN**, this clearly reveals that BERT as an excellent pre-training encoder indeed improve the richness of textual embedding, compared with LSTM-based encoder. 2) Our approach outperforms the current baselines significantly. We speculate that there are some reasons as follows: First, the cross-modal relation module devotes to refine a high-quality vision expression effectively. Second, our approach defines aspect sentiment classification as a multi-aspect task, which considerate the mutual interaction of multiple aspect sentiment.

4.4 Analysis

In this section, we give a further investigation of some experimental results and discussion of some meaningful cases.

Ablation Study. To further demonstrate the assistance of image-text relation, we remove relation separately i.e., remove all (**W/o Relation All**), remove image-to-aspect relation (**W/o Relation MATE**) and remove image-to-sentiment relation (W/o Relation MASC). Moreover, to demonstrate the importance of modeling image to our joint task, we remove the vision information, i.e., remove image-to-aspect vision (**W/o vision MATE**) and remove image-to-sentiment vision (**W/o vision MASC**). From Table 7, we observe that removing either the image vision or the image-text relation significantly decreases the performance. This illustrates the effectiveness of our approach in refining the visual information and modalities fusion assistance.

Case Study. To further demonstrate the effectiveness of our multi-modal joint task approach, Figure 3 presents three examples with predicted result by **JML**, and three representative baselines **D-GCN**, **OSCGA-collapse** and **JML w/o relation all**. We can obviously realize that: In example (a), although **D-GCN** can accurately detect two aspect terms of ground-truth, it gives the wrong sentiment prediction of aspect term "lionelmessi". This is mainly because of the lack of auxiliary visual information. In example (b), **OSCGA-collapse** predicts an error aspect, owing to incorporate collapsed tags with the same visual feeding in process of mine the visual information for these two sub-tasks. In example (c), we found that **JML w/o relation all** predict an error sentiment of aspect "miami", suggesting that without the assistance of cross-modal relation, the approach receives the interference of useless image information. However, from these cases, we observe that our well-behaved approach **JML** can obtain all correct aspect terms and aspect-dependent sentiment by controlling the inflow of image information and separately mine the visual information for two sub-tasks in a joint framework.

5 Conclusion

In this paper, we propose a multi-modal joint approach to simultaneously handle the aspect terms extraction and sentiment classification. Our approach can not only model the cross-modal relation between text and image, determining how much

Figure 3: Three cases of the predictions by **D-GCN**, **OSCGA-collapse**, **JML w/o relation all**, and **JML**. *Pos*: Positive, *Neu*: Neutral, *Neg*: Negative.

visual information contributes to text, but also separately mine the visual information for two sub-tasks instead of collapsed tagging with the same visual feeding. The detailed evaluation demonstrates that our proposed model significantly outperforms several state-of-the-art baselines.

In our future work, we will extend our approach to more multi-modal multi-task scenarios, such as relation extraction and emotion cause extraction in multi-modal dialogue. Furthermore, we would like to investigate other approaches (e.g.,self-supervised neural network) to better model JMASA.

6 Acknowledgments

We thank our anonymous reviewers for their valuable suggestions. This work was supported by the National Key R& D Program of China under Grant No. 2020AAA0108600, an NSFC grant No.62076176 and a project funded by China Postdoctoral Science Foundation No.2020M681713.

References

Md. Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of NAACL-HLT 2019*, pages 370–379. Association for Computational Linguistics.

Md. Shad Akhtar, Tarun Garg, and Asif Ekbal. 2020. Multi-task learning for aspect term extraction and aspect sentiment classification. *Neurocomputing*, 398:247–256.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of COLING 2020*, pages 272–279. International Committee on Computational Linguistics.

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of ACL 2019*, pages 547–556.

Zhuang Chen and Tieyun Qian. 2020a. Enhancing aspect term extraction with soft prototypes. In *Proceedings of EMNLP 2020*, pages 2107–2117. Association for Computational Linguistics.

Zhuang Chen and Tieyun Qian. 2020b. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of ACL 2020*, pages 3685–3694.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR 2016*, pages 770–778. IEEE Computer Society.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of ACL 2019*, pages 537–546. Association for Computational Linguistics.

Yunjie Ji, Hao Liu, Bolei He, Xinyan Xiao, Hua Wu, and Yanhua Yu. 2020. Diversified multiple instance learning for document-level multi-aspect sentiment classification. In *Proceedings of EMNLP 2020*, pages 7012–7023.

Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-based label set generation for multi-modal multi-label emotion detection. In *Proceedings of ACM MM 2020*, pages 512–520.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. In *Proceedings of EMNLP-IJCNLP 2019*, pages 4610–4620. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

Bin Liang, Rongdi Yin, Lin Gui, Jiachen Du, Yulan He, and Ruifeng Xu. 2020a. Aspect-invariant sentiment features learning: Adversarial multi-task learning for aspect-based sentiment analysis. In *Proceedings of CIKM 2020*, pages 825–834. ACM.

Bin Liang, Rongdi Yin, Lin Gui, Jiachen Du, and Ruifeng Xu. 2020b. Jointly learning aspect-focused and inter-aspect relations with graph convolutional networks for aspect sentiment analysis. In *Proceedings of COLING 2020*, pages 150–161. International Committee on Computational Linguistics.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of ACL 2018*, pages 1990–1999. Association for Computational Linguistics.

Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. 2019. DOER: dual cross-shared RNN for aspect term-polarity co-extraction. In *Proceedings of ACL 2019*, pages 591–601. Association for Computational Linguistics.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of ACL 2019*, pages 3538–3547. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of NAACL-HLT 2018*, pages 852–860. Association for Computational Linguistics.

Chengai Sun, Liangyu Lv, Gang Tian, and Tailu Liu. 2021a. Deep interactive memory network for aspect-level sentiment analysis. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(1):3:1–3:12.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021b. Rpbert: A text-image relation propagation-based BERT model for multimodal NER. *CoRR*, abs/2102.02967.

Mukuntha Narayanan Sundararaman, Zishan Ahmad, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Unsupervised aspect-level sentiment controllable style transfer. In *Proceedings of AACL/IJCNLP 2020*, pages 303–312.

Jialong Tang, Ziyao Lu, Jinsong Su, Yubin Ge, Linfeng Song, Le Sun, and Jiebo Luo. 2019. Progressive self-supervised attention learning for aspect-level sentiment analysis. In *Proceedings ofACL 2019*, pages 557–566. Association for Computational Linguistics.

Stéphan Tulkens and Andreas van Cranenburgh. 2020. Embarrassingly simple unsupervised aspect extraction. In *Proceedings of ACL 2020*, pages 3182–3187. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS 2017*, pages 5998–6008.

Alakananda Vempala and Daniel Preotiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of ACL 2019*, pages 2830–2840.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of AAAI 2020*, pages 9122–9129.

Shuohang Wang and Jing Jiang. 2017. Machine comprehension using match-lstm and answer pointer. In *Proceedings of ICLR 2017*. OpenReview.net.

Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020a. Multimodal aspect extraction with region-aware alignment network. In *Proceedings of NLPCC 2020*, volume 12430 of *Lecture Notes in Computer Science*, pages 145–156. Springer.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020b. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of ACM MM 2020*, pages 1038–1046. ACM.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of AAAI 2019*, pages 371–378.

Chengcan Ying, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Opinion transmission network for jointly improving aspect-oriented opinion words extraction and sentiment classification. In *Proceedings of NLPCC 2020*, pages 629–640.

Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of IJCAI 2019*, pages 5408–5414. ijcai.org.

Jianfei Yu, Jing Jiang, and Rui Xia. 2020a. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:429–439.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020b. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of ACL 2020*, pages 3342–3352. Association for Computational Linguistics.

AmirAli Bagher Zadeh, Yansheng Cao, Smon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2020. CMU-MOSEAS: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of EMNLP 2020*, pages 1801–1812. Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, Dawei Song, and Benyou Wang. 2020a. A multi-task learning framework for opinion triplet extraction. In *Proceedings of EMNLP 2020*, pages 819–828. Association for Computational Linguistics.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020b. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of EMNLP 2021 2020*, pages 3584–3593. Association for Computational Linguistics.

Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021a. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *Proceedings of AAAI 2021*, pages 14338–14346. AAAI Press.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021b. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of AAAI 2021*, pages 14347–14355.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of EMNLP 2020*, pages 3540–3549.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of AAAI 2018*, pages 5674–5681. AAAI Press.

Yaowei Zheng, Richong Zhang, Samuel Mensah, and Yongyi Mao. 2020. Replicate, walk, and stop on syntax: An effective neural network model for aspect-level sentiment classification. In *Proceedings of AAAI 2020*, pages 9685–9692.