Joint Multimodal Aspect Sentiment Analysis with Aspect Enhancement and Syntactic Adaptive Learning

Linlin Zhu , Heli Sun * , Qunshu Gao , Tingzhou Yi , Liang He
Xi'an Jiaotong University
{zhulinlin, jiubaoyibao}@stu.xjtu.edu.cn, 13673559526@163.com, {hlsun, lhe}@xjtu.edu.cn

Abstract

As an important task in sentiment analysis, joint multimodal aspect sentiment analysis (JMASA) has received increasing attention in recent years. However, previous approaches either i) directly fuse multimodal data without fully exploiting the correlation between multimodal input data, or ii) equally utilize the dependencies of words in the text for sentiment analysis, ignoring the differences in the importance of different words. To address these limitations, we propose a joint multimodal sentiment analysis method based on **A**spect **E**nhancement and **S**yntactic **A**daptive **L**earning (**AESAL**). Specifically, we construct an aspect enhancement pre-training task to enable the model to fully learn the correlation of aspects between multimodal input data. In order to capture the differences in the importance of different words in the text, we design a syntactic adaptive learning mechanism. First, we construct different syntactic dependency graphs based on the distance between words to learn global and local information in the text. Second, we use a multi-channel adaptive graph convolutional network to maintain the uniqueness of each modality while fusing the correlations between different modalities. Experimental results on benchmark datasets show that our method outperforms state-of-the-art methods.

1 Introduction

In the present context of information society, multimodal data plays a pivotal role in understanding human sentiments [Zhang *et al.*, 2022a; Zhou *et al.*, 2021; Yu *et al.*, 2022]. Multimodal Aspect-Based Sentiment Analysis (MABSA), an essential AI research area, combines text, image, audio, or other modalities to accurately identify and interpret sentiments, transcending the limitations of single-modal analysis and offering a comprehensive reflection of complex sentiment expressions [Xu *et al.*, 2019; Zhu *et al.*, 2023; Das and Singh, 2023]. The proliferation of social media and online education has broadened the prospects and significance of multimodal aspect-based sentiment analysis research and application [Yang *et al.*, 2022][Wang *et al.*, 2023].

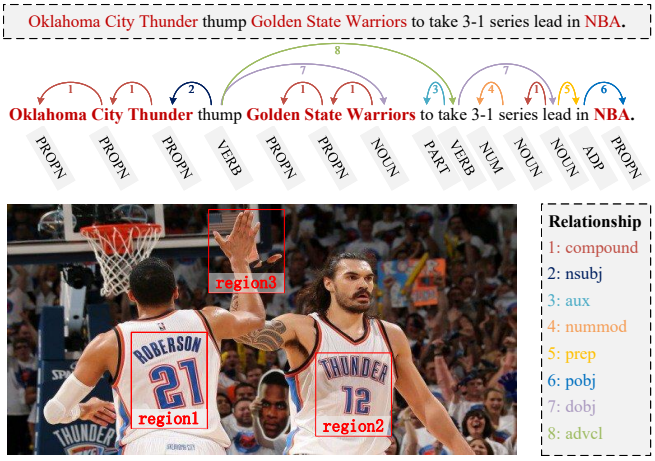

Figure 1: An example of image-text pair with its dependency graph where aspects (highlighted in red) are connected with other words based on their syntactic dependencies (different colored arrows).

Multimodal aspect sentiment analysis is usually divided into three subtasks [Yang *et al.*, 2023]: Multimodal Aspect Term Extraction (MATE), Multimodal Aspect Sentiment Classification (MASC) and Joint Multimodal Aspect Sentiment Analysis (JMASA). Earlier work on aspect-based sentiment analysis [Ju *et al.*, 2021][Ling *et al.*, 2022] directly utilized multimodal information for sentiment analysis, but did not adequately explore the complementary nature of aspects in multimodal data. The current main idea of Aspect-Based Sentiment Analysis (ABSA) is to use dependencies between sentences for modeling [Zhang *et al.*, 2022b; Tian *et al.*, 2021; Zhao *et al.*, 2020; Wang *et al.*, 2020]. Among them, [Zhang *et al.*, 2022b] proposed a Syntactically and Semantically Enhanced Graph Convolutional Network (SSEGCN) for ABSA task. [Tian *et al.*, 2021] explicitly utilized the dependency type of ABSA to categorize sentiment. These studies consider the correlation of neighboring nodes, but ignore the problem that non-neighboring nodes have different correlations.

Through the examination of the aforementioned work, we

*Contact Author

have identified two main challenges in multimodal sentiment analysis. **1) How to adequately consider aspect relevance between multimodal input data?** As shown in Fig. 1, the aspect *"Oklahoma City Thunder"* in the text is consistent with "region2" in the image, and "region3" in the image is the celebratory gesture of *"thump"* in the text. This suggests that the sentiment described in text can be reinforced or modified by the object in the image. By analyzing their interconnections, the sentiment or characteristics of the aspect can be captured and understood more accurately. **2) How to effectively capture the syntactic structure and semantic association between different words in the text?** The arrows in Fig. 1 show the dependencies of different nodes. Capturing semantic relationships between different nodes helps to understand text semantics and perform sentiment analysis.

To address the above issues, we propose a joint multimodal aspect sentiment analysis method based on aspect enhancement and syntactic adaptive learning, which fully considers the syntactic and semantic relationships between aspects and different nodes. First, we utilize RoBERTa [Liu *et al.*, 2019] and ViT [Dosovitskiy *et al.*, 2020] to encode text and image respectively. And we propose an aspect enhancement pre-training task to enable the model pay more attention to the aspect in text and image. Secondly, in order to fully consider the local and global information between nodes, we construct the syntactic dependency graphs of first-order nodes, second-order nodes and even global syntactic structure. In addition, we design a multi-channel adaptive graph convolutional network to mine the interaction features between different dependency graphs and multimodalities. Finally, the detector is used to predict the results of JMASA, MATE and MASC tasks. Our main contributions are as follows:

- We propose an AESAL model that can fully consider aspects and the syntactic associations between different nodes for JMASA, MATE and MASC tasks.
- We introduce an aspect enhancement pre-training task that constructs aspect-based positive and negative examples to enhance the sensitivity of the model to aspect term in multimodal data.
- We design the syntactic adaptive learning mechanism, which consists of syntactic dependency graph and multi-channel adaptive graph convolutional network. The syntactic dependency graphs based on different syntactic distances are used to synthesize global and local associations. The multi-channel adaptive graph convolutional network is used to enhance the node representation of the dependency graph.
- Experimental results on two benchmark datasets show that the AESAL model achieves the state of-the-art performance.

2 Related Work

In the field of multimodal sentiment analysis, some of the pre-training efforts to study specific tasks have achieved impressive results. For example, [Liu *et al.*, 2023a] introduce a unified alignment pre-training framework into the vanilla pretrain-finetune pipeline, that has both instance and knowledge-level alignment. [Ling *et al.*, 2022] propose a task-specific Vision-Language Pre-training framework for MABSA (VLP-MABSA), which is a unified multimodal encoder-decoder architecture for all the pre-training and downstream tasks. [Liu *et al.*, 2023b] propose an entity-related unsupervised pre-training with visual prompts for sentiment analysis. [Li *et al.*, 2021b] adopt supervised contrastive pre-training on large-scale sentiment annotated corpora retrieved from in-domain language resources. Unlike the above work, we design aspect enhancement pre-training task to improve the learning ability of the model for aspect term.

Recently, various approaches have been proposed to model the semantic relations between aspects and their contexts to capture opinion expressions. [Zhang *et al.*, 2022b] propose SSEGCN, which can not only learn the aspect-related semantic correlation, but also learn the global semantics of the sentence. [Liang *et al.*, 2022] give a graph convolutional network construction scheme for graph based on dependency tree and affective commonsense knowledge. [Tian *et al.*, 2021] offer a method for explicitly utilizing dependency types. [Li *et al.*, 2021a] provide a dual graph convolutional network (DualGCN) model that considers the complementarity of syntax structures and semantic correlations simultaneously. However, these approaches usually ignore the relationship between syntactic distance and syntactic semantics. To address this issue, we construct syntactic distance-based syntactic dependency graph to complement the semantic information of sentences.

Based on this, several works have extended the GCN and GAT models with syntactic dependency tree and developed several excellent models based on multimodal sentiment analysis [Zhang *et al.*, 2022b; Tian *et al.*, 2021; Zhao *et al.*, 2020; Chen *et al.*, 2020]. [Zhou *et al.*, 2023] propose an Aspect-oriented Method (AoM) which can detect aspect relevant information from perspectives of both semantics and sentiment. [Yang *et al.*, 2022] design a multitask learning architecture named Cross-Modal Multitask Transformer (CMMT) for the End-to-End sentiment analysis. [Ju *et al.*, 2021] raise a multi-modal joint learning approach with auxiliary cross-modal relation detection for multi-modal aspect-level sentiment analysis. [Ling *et al.*, 2022] present a unified multimodal encoder-decoder architecture for MABSA. Inspired by the above work, we design multi-channel adaptive graph convolutional network for multimodal sentiment analysis to maintain the uniqueness of each modality while fusing the correlations between different modalities.

In this paper, we propose a joint multimodal method to sentiment analysis based on aspect enhancement pre-training and syntactic adaptive learning. It enables the model to fully consider the aspect terms in the input, while effectively utilizing the dependencies between syntaxes.

3 Methodology

The overview of AESAL is given in Fig. 2. In this section, we describe the AESAL model, which consists of five main components: feature extraction, aspect enhancement pre-training, syntactic dependency graph, multi-channel adaptive graph

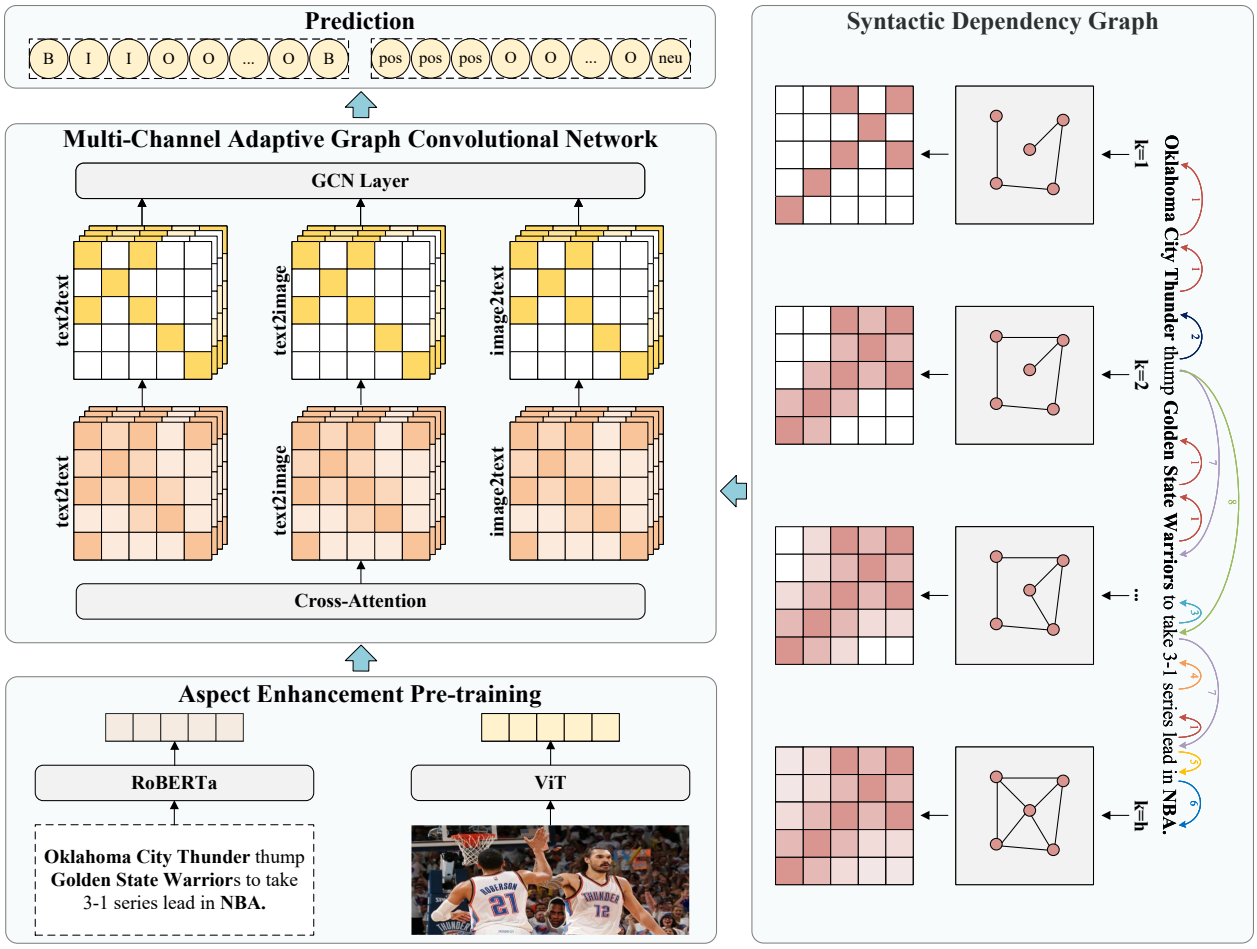

Figure 2: The overview of our proposed model AESAL.

convolutional network and prediction. Next, the components of AESAL are described separately in the remaining sections.

3.1 Feature Extractor

The multimodal data contains text sequence $T = \{t_1, t_2, ..., t_n\}$ and image V. Our goal is to extract all aspect terms $A = \{a_1, a_2, ..., a_m\}$ with the corresponding sentiment polarity $S = \{s_1, s_2, ..., s_m\}$, where n represents the length of text, m represents the number of aspects contained in the text, a_i represents the ith aspect term, and s_i represents the sentiment polarity corresponding to the ith aspect term. In addition, $s_i \in \{POS, NEU, NEG\}$, where POS, NEU and NEG standing for the sentiment of *positive*, *neutral* and *negative*, respectively.

We use RoBERTa [Yu and Jiang, 2019] as text encoder to extract the hidden context representation $H^t = \{h_1^t, h_2^t, ..., h_n^t\}$, and utilize ViT [Hu *et al.*, 2019] as image encoder to extract the hidden image representation $H^v = \{h_1^v, h_2^v, ..., h_n^v\}$. In addition, the image encoding needs to realize the alignment with the text shape by MLP. Here, H^t and $H^v \in \mathbb{R}^{n \times d}$, d represents the hidden state dimension.

3.2 Aspect Enhancement Pre-training

In order to enhance the sensitivity of the model to aspect while extracting features, we designed the aspect enhancement pre-training task, as Fig. 3. For any sample containing (T, V, A, S) quaternion, we use the special symbol "$[Mask]$" to replace the aspect A in the text sequence T to get the aspect-free text sequence N, as shown in Equation (1). We consider (T, V) as positive samples and (N, V) as negative samples to construct a new dataset (S, V, L) to improve the model learning aspect. Here, $S \in \{T, N\}$, S stands for texts, V stands for images, L stands for positive and negative sample labels, and $L = 1$ means that the aspect terms in the image match the text, otherwise 0.

$$N = T.Replace(A, [Mask]) \tag{1}$$

In the pre-training period, we first use the feature extraction module to obtain text embedding and image embedding. Then, the text embedding and image embedding are concatenated, and the probability distribution associated with the text and image is output through the softmax layer. Finally, we use cross-entropy loss to train our aspect enhancement pre-training task.

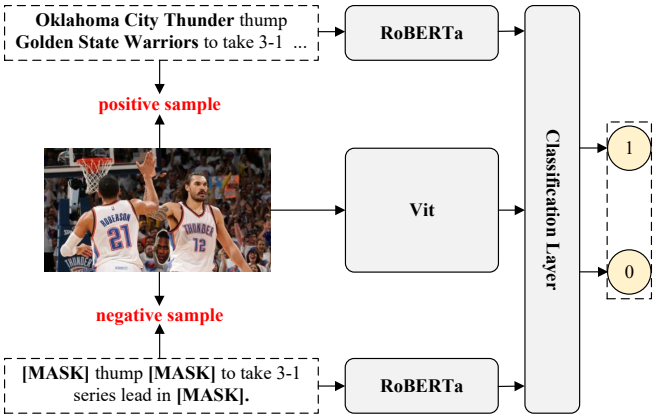

Figure 3: The framework of aspect enhancement pre-training.

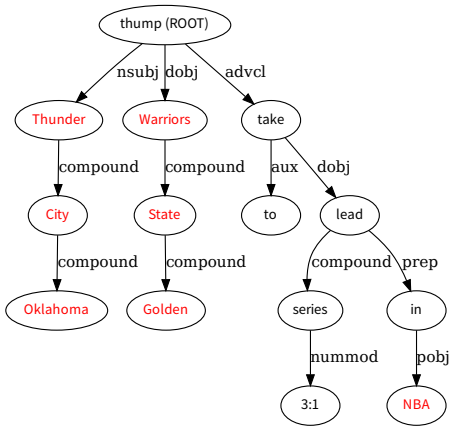

Figure 4: The dependency tree of the example mentioned in the introduction.

3.3 Syntactic Dependency Graph

Traditional syntactic dependency graph are constructed according to the connection relationship of each node in the syntactic dependency tree, which focuses only on local relationship and does not consider the deeper relationship of syntax. In order to effectively utilize the global and local dependencies between different nodes, we construct the dependency graphs based on syntactic distance, which consider both syntactic direct and indirect associations.

In this section, we use spacy[1] to obtain syntactic dependency tree, as shown in Fig. 4. A syntactic dependency tree is considered as an undirected graph, where nodes represent words and edges represent the existence of a direct relationship between two words in the syntactic structure. First, we define the distance between any two nodes i and j in the graph as $d(i,j)$. Assuming that there are a total of p paths between nodes i and j, we use the Breadth First Search (BFS) algorithm to find the shortest path d_{min} of nodes i and j, as in Equation (2). Secondly, We believe that the relationship between two distant nodes is relatively weak and there is no need for aggregation. Thus we filter the information with syntactic distance greater than k, and keep the information with syntactic distance less than or equal to k. Thus, we give more weight to the edges between two nodes with closer distance and get the dependency matrix M^k, as shown in Equation (3).

$$d_{min}\left(i,j\right) = min\left(\sum_{x=1}^{p} d_x\left(i,j\right)\right) \tag{2}$$

$$M_{i,j}^{k} = \begin{cases} \frac{k}{d_{min}(i,j)}, & d_{min}(i,j) \leq k \\ 0, & d_{min}(i,j) > k \end{cases} \tag{3}$$

Considering the different perceptual scopes at different syntactic distances, we set different thresholds for multiple (h) dependency matrices to focus on the global correlation information of the sentence and different degrees of local correlation information in Equation (4).

$$M = \left\{M^1, M^2, \ldots, M^h\right\} \tag{4}$$

[1]https://spacy.io/

M^1 captures first-order nodes dependencies, M^2 captures second-order nodes dependencies, and M^h captures global nodes dependencies.

3.4 Multi-Channel Adaptive Graph Convolutional Network

To preserve uniqueness and capture correlations across modalities and syntactic distances, we design a multi-channel adaptive graph convolutional network. It adaptively adjusts the adjacency matrix based on syntactic dependency graphs at different distances (Sec. 3.3), updating multimodal information in parallel to focus more on aspects and viewpoints.

We use the multi-head cross-attention mechanism to construct the text-to-text multi-head attention matrix P^{t2t}, text-to-image multi-head attention matrix P^{t2v}, and image-to-text multi-head attention matrix P^{v2t} respectively, and obtain the multi-channel attention matrix $P = (P^{t2t}, P^{t2v}, P^{v2t})$. It aims to capture the interrelationships that exist among text-to-text, image-to-image, and image-to-text in a multidimensional way.

First, we define the attention matrix function AF as shown in Equation (5). We define the multi-head attention matrix function MAF in Equation (6), and get the multi-channel attention matrix P in Equation (7) to match the multi-head dependency matrix M.

$$AF(\mathrm{Q},K) = Softmax\left(\frac{QW^Q \times \left(KW^K\right)^T}{\sqrt{d_k}}\right) \tag{5}$$

$$MAF(Q,K) = \left\{AF^1(Q,K), AF^2(Q,K), \ldots, AF^h(Q,K)\right\} \tag{6}$$

$$P = \left\{MAF\left(H^t,H^t\right), MAF\left(H^t,H^v\right), MAF\left(H^v,H^t\right)\right\} \tag{7}$$

Here, Q and K represent query vector and value vector respectively. W^Q and W^K are learnable parameters, d_k represents the dimension of K.

Second, we use the multi-head dependency matrix M to mask the multi-channel attention matrix P with different syntactic distances to obtain the neighbor matrix $A = (A^{t2t}, A^{t2v}, A^{v2t})$ as shown in Equation (8). At this point,

our adjacency matrix A contains syntactic dependencies at different distances and contains learnable parameters. Thus, we can both realize multi-level information aggregation from local to whole at the syntactic level and adaptively adjust the filtering of aspect-independent information.

$$A = M \odot \left\{P^{t2t}, P^{t2v}, P^{v2t}\right\} \tag{8}$$

where $\odot$ denotes element-wise multiplication.

Finally, we feed the text-hidden state H^t, image-hidden state H^v with the adjacency matrix A into the graph convolutional network to obtain features H^{gt2t}, H^{gt2v} and H^{gv2t} as in Equation (9). And the convolutional features H^{gt2t}, H^{gt2v}, H^{gv2t} take the mean value as our multimodal fusion feature H^{fusion}, as in Equation (10).

$$\begin{aligned} H^{gt2t} &= ReLU\left(A^{t2t}H^tW^{t2t}\right) \\ H^{gt2v} &= ReLU\left(A^{t2v}H^vW^{t2v}\right) \\ H^{gv2t} &= ReLU\left(A^{v2t}H^tW^{v2t}\right) \end{aligned} \tag{9}$$

$$H^{fusion} = mean\left(H^{gt2t} + H^{gt2v} + H^{gv2t}\right) \tag{10}$$

3.5 Prediction

We use a decoder consisting of a two-layer linear neural network with an activation function for task prediction, and a loss function using cross-entropy loss with the following equation:

$$\hat{y} = Softmax\left(ReLU\left(H^{fusion}W_1 + b_1\right)W_2 + b_2\right) \tag{11}$$

$$\mathcal{L} = -\sum_{i=1}^{m} y_i \log\left(\hat{y}_i\right) \tag{12}$$

where W_1, W_2, b_1, b_2 are the learnable parameters, $\hat{y}$ is the subtask prediction result, y is the subtask true label, $\mathcal{L}$ is the final loss.

4 Experiment

We compare our model with numerous methods on three tasks, including JMASA, MATE and MASC.

4.1 Experimental Settings

Datasets. We conduct experiments on two public Twitter datasets [Yu and Jiang, 2019] (i.e., Twitter-2015 and Twitter-2017). As shown in Table 1, sentences with multiple aspects make up a significant portion of both datasets. We use these datasets for aspect enhancement pre-training and subsequent experiments.

Implementation Details. We implement our method under Linux system, CUDA version 10.2, Pytorch version 1.12.0, Python version 3.9, and NVIDIA GeForce RTX 3090. In addition, we set the Learning rate to 2e-5, the dropout to 0.1, hidden size to 768.

Evaluation Metrics. We evaluate the performance of our model on JMASA task and MATE task by Micro-F1 score (F1), Precision (P) and Recall (R), while on MASC task we use Accuracy (Acc) and F1 following previous studies.

	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1,508	515	493
Neutral	1,883	670	607	1,638	517	573
Negative	368	149	113	416	144	168
# one aspect	2,159(61.65%)			976(33.54%)		
# mult. aspects	1,343(38.35%)			1,934(66.46%)		
Total Aspects	3,502			2,910		

Table 1: The basic statistics of two Twitter datasets. "# X" indicates the number of X. "mult. aspects" stands for "multiple aspects".

4.2 Baselines

We compare AESAL with four types of methods listed below.

Methods for textual ABSA. 1) **SPAN** [Hu *et al.*, 2019] is a span-based extraction-categorization framework that extracts multiple opinions directly from sentences and then categorizes them. 2) **D-GCN** [Chen *et al.*, 2020] performs the task following the sequence tagging paradigm and models the dependencies between input words with an appropriate architecture. 3) **BART** [Yan *et al.*, 2021] is a pre-trained model to solve seven ABSA subtasks.

Methods for JMASA. 1) **UMT-collapse** [Yu *et al.*, 2020], **OSCGA-collapse** [Wu *et al.*, 2020b] and **RpBERT-collapse** [Sun *et al.*, 2021] use the same visual feed to collapse the markers. 2) **UMT+TomBERT, OSCGA+TomBERT** are two pipeline methods. 3) **JML** [Ju *et al.*, 2021] is a multimodal joint approach to simultaneously handle the aspect terms extraction and sentiment classification. 4) **VLP-MABSA** [Ling *et al.*, 2022] is a unified multimodal encoder-decoder architecture for all the pre-training and downstream tasks. 5) **CMMT** [Yang *et al.*, 2022] is a multi-task learning framework to extract aspect-sentiment pairs from a pair of sentence and image. 6) **AOM** [Zhou *et al.*, 2023] is an aspect-oriented network to mitigate the visual and textual noises from the complex image-text interaction.

Methods for MATE. 1) **RAN** [Wu *et al.*, 2020a] is a novel approach which uses object and text features as the input on MATE task. 2) **UMT** [Yu *et al.*, 2020] is a unified architecture to alleviate the bias of the visual context in multimodal named entity recognition. 3) **OS-CGA** [Yan *et al.*, 2021] is an object-aware neural model that combines visual and textual representations into entities predicting.

Methods for MASC. 1) **ESAFN** [Yu *et al.*, 2019] is an entity-sensitive attention and fusion network for MASC. 2) **TomBERT** [Yu and Jiang, 2019] is a target-oriented multimodal sentiment classification method. 3) **CapTrBERT** [Khan and Fu, 2021] is a two-stream model that translates images in input space.

4.3 Main Results

In this section, we show the excellent performance of our method compared with SOTAs.

Performance on JMASA. The results of JMASA are shown in Table 2. First, our model far outperforms all text-based models, which means that utilizing multimodal information

	Methods	Twitter-2015			Twitter-2017		
		P	R	F1	P	R	F1
Text-based	SPAN*	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN*	58.3	58.8	59.4	64.2	64.1	64.1
	BART*	62.9	65.0	63.9	65.2	65.6	65.4
Multimodal	UMT+TomBERT*	58.4	61.3	59.8	62.3	62.4	62.4
	OSCGA+TomBERT*	61.7	63.4	62.5	63.4	64.0	63.7
	OSCGA-collapse*	63.1	63.7	63.2	63.5	63.5	63.5
	RpBERT-collapse*	49.3	46.9	48.0	57.0	55.4	56.2
	UMT-collapse*	61.0	60.4	61.6	60.8	60.0	61.7
	JML	65.0	63.2	64.1	66.5	65.5	66.0
	VLP-MABSA	65.1	68.3	66.6	66.9	69.2	68.0
	CMMT	64.6	68.7	66.5	67.6	69.4	68.5
	AoM	67.9	69.3	68.6	68.4	71.0	69.7
	AESAL(ours)	**68.7**	**70.4**	**69.5**	**69.4**	**74.8**	**72.0**

Table 2: Results of different methods for JMASA on the two Twitter datasets. Our model AESAL achieves the current optimal results on JMASA. * denotes the results from [Zhou *et al.*, 2023]. The best results are bold-typed and the second best ones are underlined.

Methods	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
RAN*	80.5	81.5	81.0	90.7	90.7	90.0
UMT*	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA*	81.7	82.1	81.9	90.2	90.7	90.4
JML	83.6	81.2	82.4	92.0	90.7	91.4
VLP-MABSA	83.6	87.9	85.7	90.8	92.6	91.7
CMMT	83.9	88.1	85.9	92.2	93.9	93.1
AoM	84.6	87.9	86.2	91.8	92.8	92.3
AESAL(ours)	**90.2**	**90.6**	**90.4**	**93.1**	**96.4**	**94.7**

Table 3: Results of different methods for MATE. Our model AESAL achieves the current optimal results on MATE. * denotes the results from [Zhou *et al.*, 2023].

Methods	Twitter-2015		Twitter-2017	
	Acc	F1	Acc	F1
ESAFN*	73.4	67.4	67.8	64.2
TomBERT*	77.2	71.8	70.5	68.0
CapTrBERT*	78.0	73.2	72.3	70.2
JML	78.7	-	72.7	-
VLP-MABSA	78.6	73.8	73.8	71.8
CMMT	77.9	-	73.8	-
AoM	**80.2**	**75.9**	76.4	75.0
AESAL(ours)	80.1	75.2	**78.8**	**75.9**

Table 4: Results of different methods for MASC. Our model AESAL achieves the current optimum of MASC on a slightly larger dataset. * denotes the results from [Zhou *et al.*, 2023].

is beneficial for the JMASA task. Second, our model outperforms other multimodal aspect sentiment analysis methods on every metric. In particular, compared to the suboptimal model (AoM), the precision improves by 0.8%, recall improves by 1.1%, and F1 score improves by 0.9% on the Twitter-2015 dataset; on Twitter-2017, P, R, and F1 increased by 1%, 3.8%, and 2.3%, respectively. This demonstrates the effectiveness of our model on the JMASA task.

Performance on MATE. Table 3 shows the performance of MATE. Our model still achieves optimal results. Specifically, on the Twitter-2015 dataset, compared to the suboptimal result on AoM, P improves by 5.6%, R improves by 2.7%, and F1 improves by 4.2%; on the Twitter-2017 dataset, compared to the suboptimal result on CMMT, P improves by 0.9%, R improves by 2.5%, and F1 improves by 1.6%. It proves that AESAL is the most capable of detecting aspect term from image and text.

Performance on MASC. As shown in Table 4, AESAL performs best on Twitter-2017, with 2.4% improvement in Acc and 0.9% improvement in F1 compared to the second best AoM. AoM performs slightly better than us on Twitter-2015, outperforming us by 0.1% on Acc and 0.7% on F1. This may be related to the number of aspect terms in the dataset. There are fewer aspect terms in the Twitter-2015 data than in Twitter-2017, so our aspect-enhanced pre-training may not have been utilized to its maximum capacity.

4.4 Ablation Study

In this section, we compare the variants of the AESAL in terms of the following five components to demonstrate the effectiveness of AESAL framework: Image (Img), Aspect Enhancement Pre-training (AE), Indirect Relation Dependency

Methods	JMASA						MATE						MASC			
	Twitter-2015			Twitter-2017			Twitter-2015			Twitter-2017			Twitter-2015		Twitter-2017	
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc	F1	Acc	F1
Full	**68.7**	**70.4**	**69.5**	69.4	**74.8**	**72.0**	**90.2**	**90.6**	**90.4**	**93.1**	96.4	**94.7**	**80.1**	**75.2**	**78.8**	**75.9**
w/o Img	64.1	69.0	66.5	**70.4**	73.5	71.9	88.7	84.9	86.8	87.9	**96.9**	92.2	76.0	68.8	77.6	74.3
w/o AE	*63.9*	*68.0*	*65.9*	64.0	68.1	66.0	89.6	90.2	89.9	90.5	**96.9**	93.6	76.8	72.3	77.6	72.3
w/o IR	65.7	68.8	67.2	*62.8*	*62.2*	*62.5*	88.9	88.7	88.8	*84.7*	*92.7*	*88.5*	74.2	70.5	74.3	72.8
w/o SAL	65.8	69.2	67.5	69.6	69.8	69.7	*87.4*	*84.6*	*86.0*	91.6	96.2	93.8	76.3	70.1	74.5	72.2
w/o T2T	67.5	69.1	68.3	67.4	63.2	65.2	89.6	89.7	89.6	91.4	96.0	93.6	*74.3*	*68.2*	*73.5*	*71.1*

Table 5: The performance comparison of our full model and its ablated methods on JMASA, MATE and MASC. Data in italics indicate the worst results.

Graph (IR), Syntactic Adaptive Learning (SAL), Text2Text Convolution Channel (T2T).

W/o Img is a variant of AESAL that removes image information from multimodal data and utilizes only text information.

W/o AE is a variant of AESAL without aspect enhancement pre-training module.

W/o IR is a variant of AESAL without the syntactic dependency graph of indirect relations, using only the syntactic dependency graph of neighboring nodes.

W/o SAL is a variant of AESAL without syntactic adaptive learning and utilizes only cross-attention to fuse multimodal information for prediction.

W/o T2T is a variant of AESAL without the convolution channel of Text2Text and considers only image-text, text-image related channels.

The results of the ablation experiments for the JMASA, MATE and MASC tasks are given in Table 5. Based on the JMASA task, the variant without AE and the variant without IR have the worst results. This shows the importance of the aspect enhancement pre-training module and the syntactic dependency graph of indirect relation in our model. Based on the MATE task, the variant without SAL and the variant without IR have the worst results. It thus illustrates the importance of syntactic adaptive learning and indirect relations between words for the MATE task, and they help the model to better understand and learn syntactic dependencies. Based on the MASC task, the variant without the convolution channel of Text2Text has the lowest results, reflecting the importance of textual information in the MASC task. Generally speaking, textual information expresses sentiments more accurately, while images only serve an auxiliary role.

4.5 Case Study

To further demonstrate the effectiveness of AESAL, we present two test examples with predictions from different methods. As shown in Fig. 5, for 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 may be due to the lack of syntactic adaptive learning mechanism that fail to adequately capture syntactic and semantic information. However, OSCGA-collapse and our AESAL methods correctly predict the aspects and its corresponding polarity. For example, in (b), OSCGA-collapse incorrectly detects the aspect as *"ridgetown_dhs tractor day"*, but the correct aspect term is *"ridgetown_dhs"*. This is mainly because of the lack of aspect enhancement pre-training task. In both cases, our AESAL model correctly extracts all aspects and categorizes sentiments, demonstrating the excellence of AESAL on the MATE, MASC, and JMASA tasks.

Image		
Text	# **LionelMessi** ' s bride # **antonellaRoccuzzo** ' first lady of football ' |	@ **DeltaPowerEquip** leading the parade at @ **ridgetown_dhs** tractor day !
D-GCN	(LionelMessi, Neu) (√,×) (antonellaRoccuzzo, Pos) (√,√)	(DeltaPowerEquip, Pos) (√,√) (ridgetown_dhs, Neu) (√,√)
OSCGA -collapse	(LionelMessi, Pos) (√,√) (antonellaRoccuzzo, Pos) (√,√)	(DeltaPowerEquip, Pos) (√,√) (ridgetown_dhs tractor day, Neu) (×,√)
AESAL	(LionelMessi, Pos) (√,√) (antonellaRoccuzzo, Pos) (√,√)	(DeltaPowerEquip, Pos) (√,√) (ridgetown_dhs, Neu) (√,√)

Figure 5: Two cases of the predictions by D-GCN, OSCGA-collapse and our AESAL. Pos: Positive, Neu: Neutral, Neg: Negative.

5 Conclusion

In this paper, we propose a joint multimodal aspect sentiment analysis method based on aspect enhancement and syntactic adaptive learning. First, we construct the aspect enhancement pre-training task to enable the model adequately learn the aspect of the multimodal input data. Second, we design different syntactic dependency graphs of first-order nodes, second-orders nodes and even global nodes simultaneously, so that the model captures the global and local information in the text. After that, we enhance the node representation by utilizing the multi-channel adaptive graph convolutional network. Finally, we execute the MATE, MASC, and JMASA tasks, respectively. Experimental results on two widely available datasets demonstrate the effectiveness of our method.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant 62072365, in part by the Science Foundation of Distinguished Young Scholars of Shaanxi under Grant 2022JC-48, in part by the Aviation Science Foundation 2023M071070002, in part by the Key Research and Development Program of Shaanxi under Grant 2022GY-332, 2023-YBGY-230 and 2024GX-YBXM-533, in part by the Innovation Capability Support Plan of Shaanxi under Grant 2022PT-33, in part by the Xi'an Science and Technology plan Key industrial chain technology research project under Grant 23ZDCYJSGG0007, and in part by the Xi'an Science and Technology plan Key industrial chain, key core technology research project under Grant 23LLRH0022, and Qinchuangyuan Construction of Two Chain Integration Important Project 23LLRHZDZX0006. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[Chen *et al.*, 2020] Guimin Chen, Yuanhe Tian, and Yan Song. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th international conference on computational linguistics*, pages 272–279, 2020.

[Das and Singh, 2023] Ringki Das and Thoudam Doren Singh. Multimodal sentiment analysis: A survey of methods, trends and challenges. *ACM Computing Surveys*, 2023.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Hu *et al.*, 2019] Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. Open-domain targeted sentiment analysis via span-based extraction and classification. *arXiv preprint arXiv:1906.03820*, 2019.

[Ju *et al.*, 2021] Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4395–4405, 2021.

[Khan and Fu, 2021] Zaid Khan and Yun Fu. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3034–3042, 2021.

[Li *et al.*, 2021a] Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329, 2021.

[Li *et al.*, 2021b] Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. *arXiv preprint arXiv:2111.02194*, 2021.

[Liang *et al.*, 2022] Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235:107643, 2022.

[Ling *et al.*, 2022] Yan Ling, Jianfei Yu, and Rui Xia. Vision-language pre-training for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2204.07955*, 2022.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2023a] Juhua Liu, Qihuang Zhong, Liang Ding, Hua Jin, Bo Du, and Dacheng Tao. Unified instance and knowledge alignment pretraining for aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2629–2642, 2023.

[Liu *et al.*, 2023b] Kuanghong Liu, Jin Wang, and Xuejie Zhang. Entity-related unsupervised pretraining with visual prompts for multimodal aspect-based sentiment analysis. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 481–493. Springer, 2023.

[Sun *et al.*, 2021] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868, 2021.

[Tian *et al.*, 2021] Yuanhe Tian, Guimin Chen, and Yan Song. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 2910–2922, 2021.

[Wang *et al.*, 2020] Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*, 2020.

[Wang *et al.*, 2023] Di Wang, Changning Tian, Xiao Liang, Lin Zhao, Lihuo He, and Quan Wang. Dual-perspective fusion network for aspect-based multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 2023.

[Wu *et al.*, 2020a] Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. Multimodal aspect extraction with region-aware alignment network. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9*, pages 145–156. Springer, 2020.

[Wu *et al.*, 2020b] Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1038–1046, 2020.

[Xu *et al.*, 2019] Nan Xu, Wenji Mao, and Guandan Chen. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 371–378, 2019.

[Yan *et al.*, 2021] Hang Yan, Junqi Dai, Xipeng Qiu, Zheng Zhang, et al. A unified generative framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2106.04300*, 2021.

[Yang *et al.*, 2022] Li Yang, Jin-Cheon Na, and Jianfei Yu. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing & Management*, 59(5):103038, 2022.

[Yang *et al.*, 2023] Xiaocui Yang, Shi Feng, Daling Wang, Sun Qi, Wenfang Wu, Yifei Zhang, Pengfei Hong, and Soujanya Poria. Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt. *arXiv preprint arXiv:2305.10169*, 2023.

[Yu and Jiang, 2019] Jianfei Yu and Jing Jiang. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5408–5414, 2019.

[Yu *et al.*, 2019] Jianfei Yu, Jing Jiang, and Rui Xia. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439, 2019.

[Yu *et al.*, 2020] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[Yu *et al.*, 2022] Jianfei Yu, Kai Chen, and Rui Xia. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 2022.

[Zhang *et al.*, 2022a] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[Zhang *et al.*, 2022b] Zheng Zhang, Zili Zhou, and Yanna Wang. Ssegcn: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4916–4925, 2022.

[Zhao *et al.*, 2020] Pinlong Zhao, Linlin Hou, and Ou Wu. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowledge-Based Systems*, 193:105443, 2020.

[Zhou *et al.*, 2021] Jie Zhou, Jiabao Zhao, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. Masad: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, 455:47–58, 2021.

[Zhou *et al.*, 2023] Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2306.01004*, 2023.

[Zhu *et al.*, 2023] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325, 2023.