Autonomous Aspect-Image Instruction A^{2}II: Q-Former Guided Multimodal Sentiment Classification

Junjia Feng[1,2], **Mingqian Lin**[1,2], **Lin Shang**[1,2*], **Xiaoying Gao**[3]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Department of Computer Science and Technology, Nanjing University, China
[3]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
{fengjunjia, linmingqian}@smail.nju.edu.cn, shanglin@nju.edu.cn, xgao@ecs.vuw.ac.nz

Abstract

Multimodal aspect-oriented sentiment classification (MABSC) task has garnered significant attention, which aims to identify the sentiment polarities of aspects by combining both language and vision information. However, the limited multimodal data in this task has become a big gap for the vision-language multimodal fusion. While large-scale vision-language pretrained models have been adapted to multiple tasks, their use for MABSC task is still in a nascent stage. In this work, we present an attempt to use the instruction tuning paradigm to MABSC task and leverage the ability of large vision-language models to alleviate the limitation in the fusion of textual and image modalities. To tackle the problem of potential irrelevance between aspects and images, we propose a plug-and-play selector to autonomously choose the most appropriate instruction from the instruction pool, thereby reducing the impact of irrelevant image noise on the final sentiment classification results. We conduct extensive experiments in various scenarios and our model achieves state-of-the-art performance on benchmark datasets, as well as in few-shot settings.

Keywords: sentiment analysis, vision-language, instruction tuning, twitter

1. Introduction

As one of the important subtasks of aspect-based sentiment analysis (Tang et al., 2016a; Wang et al., 2016a; Sun et al., 2019a), multimodal aspect-oriented sentiment classification (MABSC) has received more and more attention which aims to identify the sentiment polarity of the aspect in a provided sentence and image pair. Early works (Xu et al., 2019; Yu et al., 2019) are based on RNNs and designed effective attention mechanisms to model the alignment of the aspects and the image-sentence pairs. Yu and Jiang (2019) and Wang et al. (2021) followed the fine tuning paradigm and exploited pre-trained language models to compute the interactions between aspects and images together with sentences. More recently, Khan and Fu (2021) and Yang et al. (2022) tried to translate images into text modality through image captioning models, then fused the original and translated text information to predict the sentiment polarity.

However, all these existing methods suffer from two common limitations. Firstly, these works are either to treat the visual features equal to text features by employing a unimodal pre-trained language model for prediction or to translate the images into captions through a captioning model, which may cause additional loss of the image information. Meanwhile, constrained by the size of datasets and the limitation of cross-modal alignment supervision in the MABSC task, the current task-specific training corpus tends to result in sub-optimal multimodal models. With the recent trend of large language-vision models which were trained on large-scale data, it is time to exploit these models' abilities to enhance multimodal fusion for downstream tasks. However, the huge parameters hindered the coming growth. The emergence of some large language-vision models (Li et al., 2023; Dai et al., 2023) that only need to train a lightweight Querying Transformer has bridged the modality gap, which allows us to leverage the ability of modality fusion in the downstream tasks without introducing too many additional parameters. Additionally, aspects may be irrelevant to the image, which may mislead the model to output wrong predictions when we directly input irrelevant image features into the model. Some previous works considered the relationship between aspects and images, for example, Yu and Jiang (2019) models aspects and images via an aspect-image mechanism, Yu et al. (2022a) additionally introduces a manually labeled image-aspect matching dataset. Different from the previous methods, we try to solve the problem by selecting the most appropriate instruction.

Recently, instruction tuning has achieved impressive results in language models (Wei et al., 2021; Ouyang et al., 2022; Wang et al., 2022b), which enables the model to process and follow different instructions while improving their generalization performance. Although instruction tuning has been widely studied for language models, instruction tuning for multimodal tasks remains relatively less explored. In this paper, to address these two limitations and motivated by instruction learning, we

* Corresponding author.

LREC-COLING 2024, pages 1996–2005
20-25 May, 2024.

build our model following the instruction learning paradigm, and use a lightweight pre-trained Querying Transformer (Q-Former) to mitigate the problem of fusing textual and image modalities, which exploits the ability of large language-vision models, then we carefully design different instructions and then train a plug-and-play selector module to select the most appropriate instruction autonomously, i.e. when the image and aspect are related, the instruction indicates that the model needs to combine the image information as much as possible to identify the sentiment prediction; when the image and aspect are irrelevant, the instruction indicates the model to make sentiment predictions based solely on text information, so as not to be affected by irrelevant image features in the large part.

Experimental results on two benchmark datasets show that our model outperforms several state-of-the-art methods and can better fuse image and aspect information. Furthermore, experiments have been carried out in various scenarios demonstrating the efficacy of our instruction tuning framework in enhancing model robustness within image noise scenarios and its benefits in few-shot settings. Our main contributions are summarized as follows:

- Considering the limitations of the datasets, we exploit large language-vision models for cross-modal fusion without introducing too many additional parameters and explore instruction tuning for multimodal aspect-oriented sentiment classification (MABSC) task.
- To tackle the challenge of potentially irrelevant images and aspects, we introduce a selector module designed to guide the model in autonomously choosing appropriate instructions, thereby mitigating the influence of image noise.
- We conduct extensive experiments in various scenarios and our model achieves state-of-the-art performance on benchmark datasets, as well as in few-shot settings.

2. Related Work

2.1. Multi-Modal Aspect-Oriented Sentiment Classification

As social media continues to develop, multimodal data has become a prominent source of information in addition to text. Therefore, researchers proposed a novel subtask of aspect-based sentiment analysis called multimodal aspect-oriented sentiment classification, which is garnering increasing attention. Xu et al. (2019) first proposed this task and presented a Multi-Interactive Memory Network (MIMN) based on RNNs. Yu et al. (2019) presented an entity-sensitive attention fusion network with a simple gate mechanism to reduce the influence of noise in the image. With the continuous development of pre-training models, many works have been proposed based on the fine-tuning paradigm (Yu and Jiang, 2019; Khan and Fu, 2021; Wang et al., 2021; Ling et al., 2022; Yang et al., 2022). In contrast to prior research, we explore instruction tuning paradigm for MABSC task.

2.2. Instruction Tuning

instruction tuning paradigm has greatly enhanced the generalization performance of large language models (Wei et al., 2021; Ouyang et al., 2022; Wang et al., 2022b; Chung et al., 2022a; Wang et al., 2022a; Honovich et al., 2022) and has achieved good results in a variety of tasks. Wei et al. (2021) showed that instruction tuning can substantially improve zero-shot performance of language models on unseen tasks. Wang et al. (2022b) builded Tk-INSTRUCT, a transformer model trained to follow a variety of incontext instructions which outperforms a serious of instruction-following language models. While instruction tuning paradigm has been widely used for language models, there have been some recent works focusing on vision-language instruction tuning (Xu et al., 2022; Liu et al., 2023; Zhu et al., 2023; Dai et al., 2023; Ye et al., 2023).

3. Methodology

3.1. Task Definition

Given a set of multimodal samples M, for each sample $m_i \in M$, it contains a sentence $S_i = (w_1, w_2, ..., w_n)$, an image I_i corresponding to the sentence, and an aspect T_i, where n is the number of words in the sentence, and T_i is a subsequence of S_i. Each aspect T_i is annotated with a label $y_i \in \{positive, neutral, negative\}$. Therefore, the task can be defined as giving a set of multimodal samples M as the training set, and the goal is to learn a sentiment classifier such that the classifier can correctly predict the sentiment labels of corresponding aspects in unseen multimodal samples.

3.2. Overview

Figure 1 shows the overview of our model architecture. Given a multimodal input sample $m_i = (S_i, I_i, T_i)$, which contains a sentence S_i, a corresponding image I_i and an aspect T_i, we first take the image I_i into an image encoder to generate the global features of the image. Let $3, H, W$ represent the number of channels, width and height of an image respectively. The image encoder turns the image $I_i \in \mathcal{R}^{3\times H\times W}$ into a tensor $V \in \mathcal{R}^{l\times d}$, where l and d represent the length of processed image

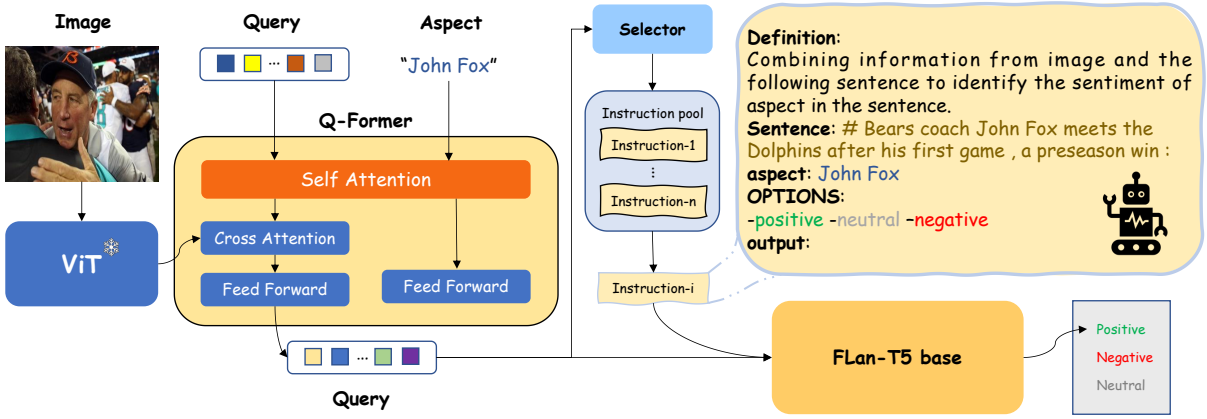

Figure 1: Overview of A^{2}II model architecture for MABSC.

and the hidden size of the image encoder respectively. In addition, since the visual input V contains the global features instead of aspect-related fine-grained features, it is necessary to match the aspect with the relevant fine-grained image features. Therefore, we adopt Q-Former module used in instructBLIP (Dai et al., 2023) to align the image and the aspect and then obtain the multimodal output F. Subsequently, since the aspect and the image may not be related, the model can autonomously select the appropriate instruction Ins_i from the instruction pool $P = (Ins_1, ..., Ins_n)$ through the selector module according to the multimodal feature. Then we join the selected instruction Ins_i and the sentence S_i together to get the instruction input E_i. Finally, we pass the multimodal feature F through a language projection layer and then concatenate it with the instruction input E, feeding them through the language model to obtain a sentiment prediction y_i for the aspect. In the following subsections, we will introduce each module in detail.

3.3. Multimodal Fusion Module

In this module, we aim to leverage the ability of large language-vision models to obtain image-aspect related multimodal features. Firstly, given the input image $I_i \in \mathcal{R}^{3\times H\times W}$, we apply a frozen image encoder ViT-g/14 (Fang et al., 2023) to generate image features $V \in \mathcal{R}^{l\times d}$, where l and d represent the length of processed image and the hidden size of the image encoder respectively.

$$ViT(I_i) = \{r_j \mid r_j \in \mathcal{R}^d, j = 1, 2, ..., l\} \quad (1)$$

Then, we create a set number of learnable query embeddings $Z \in \mathcal{R}^{32\times q}$, where q represents the hidden size of the Q-Former. These queries can interact with aspect features through self-attention layers and also engage with image features through cross-attention layers. Just as instructBLIP inputting image features and instructions into Q-Former together to make queries Z extract image features which are more informative of the task as described by the instructions, we also input image features and aspects into Q-Former at the same time to obtain image-aspect related fine-grained features. As the Q-Former has been pre-trained to extract language-informative visual representations via queries, it effectively operates as an information bottleneck, providing the most useful information while filtering out irrelevant visual information. Finally, we get the final hidden state F_h and the pooler output F_p of the queries as the multimodal fusion features:

$$F_h, F_p = QFormer(Z) \quad (2)$$

where $F_h \in \mathcal{R}^{32\times q}$, $F_p \in \mathcal{R}^q$ and $QFormer(\cdot)$ denotes the self-attention layers and cross-attention layers in Q-Former model.

3.4. Instruction Selector Module

Considering that images and aspects may not be related, we design this module to autonomously select the appropriate instruction. We feed $F_p \in \mathcal{R}^q$ to a linear function followed by a softmax function for instruction selection:

$$p(y \mid F_p) = softmax(W_S^T F_p) \quad (3)$$

where $W_S \in \mathcal{R}^{q\times n}$ is the weight matrix and n denotes the number of instructions in the instruction pool. Then, we choose $k = argmax(p(y \mid F_p))$, which denotes the k^{th} instruction in the instruction pool P. Here we set n to 2, i.e. we use two instructions which guide the language model in determining the relevance of the image and aspect. To be more specific, the two instructions are "Definition: Combining information from image and the following sentence to identify the sentiment of aspect in the sentence. " and "Definition: Based solely on the information in the following sentence to identify the sentiment of aspect in the sentence.", which

are used in image-aspect related scenarios and image-aspect unrelated scenarios respectively.

3.5. Sentiment Prediction Module

Given instruction Ins_k from the instruction selector module, we then concatenate the instruction Ins_k with sentence S_i and aspect T_i to obtain the instruction input E_i. In addition, we pass the multimodal fusion features $F_h \in \mathcal{R}^{32\times q}$ through a language projection layer and get $F_l \in \mathcal{R}^{32\times g}$, where g represent the hidden size of the language model. Subsequently, we concatenate F_l with the instruction input E_i and present the general language model with three sentiment options, allowing it to generate the correct prediction:

$$\hat{y_i} = Language\_Model(F_l, E_i) \tag{4}$$

During training phase, to optimize all the parameters in our model, the objective is to minimize the loss function as below:

$$\mathcal{L} = -\sum_{i=1}^{N} P(\hat{y_i} \mid F_l, E_i) \tag{5}$$

where $P(\hat{y_i} \mid F_l, E_i)$ is the probability predicted by the general language model.

4. Experiments

4.1. Experimental Settings

4.1.1. Datasets

We carry out experiments on two benchmark datasets for MABSC task: Twitter-2015 dataset and Twitter-2017 dataset, which respectively include multimodal user posts published during 2014-2015 and 2016-2017 on Twitter. These two datasets comprise multimodal tweets that have been annotated for the mentioned aspect within their textual content (Zhang et al., 2018; Lu et al., 2018), along with the sentiment polarity associated with each aspect (Yu and Jiang, 2019). Each multimodal tweet is constructed as a pairing of an image and a sentence, which contains the designated target aspect. An overview of key details for both datasets is provided in Table 1.

4.1.2. Implementation Details

We construct our frozen image encoder using ViT-g/14 (Fang et al., 2023). On this foundation, we develop a Q-Former model, building upon the instructBLIP model introduced by Dai et al. (2023). And our language model is constructed using FlanT5-base (Chung et al., 2022b), which is an instruction-tuned model based on the encoder-decoder Transformer T5 (Raffel et al., 2020). Then, we fine-tune hyperparameters on the development set of each dataset, maintaining the frozen state of the image encoder. The AdamW optimizer (Loshchilov and Hutter, 2017) is employed with a learning rate of 5e-5. Moreover, we set the mini-batch size to 8. For input sequences and output sequences, as well as sentiment polarities, maximum lengths are set as 150 and 5 respectively. The hidden dimension of both the Q-Former and language model is set as 768. We use the accuracy metric and Macro F1-score for MABSC Task, following previous approaches (Yang et al., 2022; Khan and Fu, 2021; Yu and Jiang, 2019). All models are implemented on Pytorch with an NVIDIA Tesla V100 GPU.

4.2. Baseline Methods

We compare our method with the following baselines on Twitter-2015 and Twitter-2017 datasets and reported the accuracy and Macro-F1 score in Table 2. (1)Res-Entity (He et al., 2016) is only based on the image information, directly uses the visual feature of the input image from ResNet. (2)AE-LSTM (Wang et al., 2016b) is based on LSTM and adds attention mechanism on aspect embeddings. (3)MemNet (Tang et al., 2016b) uses a multi-hop attention mechanism on top of word embeddings and position embeddings with aspects as queries. (4)RAM (Chen et al., 2017) is based on LSTM and considers the relative distance of aspects. (5)BERT (Sun et al., 2019b) directly uses sentence pair mode to deal with aspect-level sentiment classification problems. (6)MIMN (Xu et al., 2019) is a multi-interactive memory network to obtain the connection between images and sentences, then calculates the feature representations of images and sentences based on the recurrent neural network and combined with aspect information to obtain classification results. (7)ESAFN (Yu et al., 2019) is an entity-sensitive attention fusion network based on a recurrent neural network, which simultaneously calculates image features, text features, and multimodal features to predict classification results. In addition, a gate mechanism is introduced to reduce the influence of noise in the image on the results. (8)TomBERT (Yu and Jiang, 2019) uses ResNet to obtain image features, uses BERT model to obtain aspect and sentence features, then uses cross-attention mechanism to establish the connection between images and aspects, and finally uses self-attention layers to obtain multimodal features for sentiment classification. (9) CapTr-BERT (Khan and Fu, 2021) firstly generates image captions and constructs auxiliary sentences to convert the visual representation into the same semantic space as the text representation, and then directly uses BERT model to classify the sentiments. In addition, CapBERT-DE replaces BERT with BERTweet (Nguyen et al., 2020) on the basis of

	Twitter-15							Twitter-17						
	Pos	Neg	Neu	Total	Aspects	Words	Length	Pos	Neu	Neg	Total	Aspects	Words	Length
Train	928	368	1883	3179	1.348	9023	16.72	1508	416	1638	3562	1.410	6027	16.21
Dev	303	149	670	1122	1.336	4238	16.74	515	144	517	1176	1.439	2922	16.37
Test	317	113	607	1037	1.345	3919	17.05	493	168	573	1234	1.450	3013	16.38

Table 1: Basic statistics of two benchmark datasets for multimodal aspect-oriented sentiment classification task. Pos: Positive, Neg: Negative, Neu: Neutral

CapTrBERT. (10)VLP-MABSA (Ling et al., 2022) is a task-specific pre-training vision language model for MABSA. (11)SaliencyBERT (Wang et al., 2021) uses a recurrent attention network over the BERT architecture for MABSC task. (12)FITE (Yang et al., 2022) utilizes facial emotions from images and selectively matches and fuses with the aspect in textual modality. In addition, FITE-DE replaces BERT with BERTweet (Nguyen et al., 2020) on the basis of FITE. (13)InstructBLIP (Dai et al., 2023) is a simple yet novel instruction tuning framework which is the state-of-the-art model towards generalized large vision-language models. And we test it for the MABSC task in zero-shot setting.

4.3. Experimental Results and Analysis

We have conducted a comprehensive comparison of our methods against the above baselines, utilizing both the Twitter-2015 and Twitter-2017 datasets. Notably, the most remarkable scores for each metric have been denoted in bold. From our analyses, several key observations can be drawn. Firstly, we can find that the single-modal method based on pure images performs very poorly, which reflects that in social media scenarios, texts are of great significance for aspect-level sentiment classification task, and text information is still the most important factor to reflect user sentiments. Secondly, most of the multi-modal methods based on images and texts are better than the single-modal methods, which reflects that considering the texts and images posted by users at the same time in the social media scene is more conducive to mining the sentiment polarities of users. Furthermore, in the realm of text-based methods, it becomes evident that BERT consistently surpasses all competing baselines, which underscores the efficacy of a strong pre-trained model. This phenomenon can be attributed to the inherent capacity of pre-trained models to offer better text features. In light of this observation, it is suggested that we can use the pre-trained large vision-lanuage model in the fusion of image and text, instead of just fine-tuning the language model like bert based on the small-scale multimodal data in the specific domain for modality fusion. Finally, we find that directly using instructBLIP for MABSA task does not work very well, which may be due to the fact that it is difficult for large vision-lanuage models to recognize fine-grained sentiments in multimodal data, therefore it is essential to fine-tune large vision-lanuage models using task-specific data for MABSC task.

Method	Twitter-2015		Twitter-2017	
	Acc	Macro-F1	Acc	Macro-F1
	Image Only			
Res-Target	59.88	46.48	58.59	53.98
	Text Only			
AE-LSTM	70.30	63.43	61.67	57.97
MemNet	70.11	61.76	64.18	60.90
RAM	70.68	63.05	64.42	61.01
Bert	74.15	68.86	68.15	65.23
	Text and Image			
MIMN	71.84	65.69	65.88	62.99
ESAFN	73.38	67.37	67.83	64.22
TomBERT	77.15	71.15	70.34	68.03
CapBERT	78.01	73.25	69.77	68.42
CapBERT-DE	77.92	73.90	72.30	70.20
VLP-MABSA	78.60	73.80	73.80	71.80
SaliencyBERT	77.03	72.36	69.69	67.19
FITE	78.49	73.90	70.90	68.70
FITE-DE	78.64	74.30	72.98	71.97
InstructBLIP	57.57	59.63	60.37	35.96
A^2II (Ours)	**79.46**	**75.16**	**74.39**	**72.35**

Table 2: Experiment results for multimodal aspect-oriented sentiment classification task.

Our method performs better in these two datasets compared with all the image-only, text-only and multimodal baselines. This demonstrates the effectiveness of the proposed method. On Twitter15, the accuracy has 0.82 point improvement (78.64→79.46) and the macro F1-score has 0.86 point improvement (74.30→75.16) compared with the previous best methods FITE-DE. On Twitter17, the accuracy has 0.59 point improvement (73.80→74.39) and the macro F1-score has 0.38 point improvement (71.97→72.35) compared with the previous best methods VLP-MABSA and FITE-DE. In addition, VLP-MABSA is a complex task-specific pre-trained multimodal model for MABSA with three types of pre-training tasks: textual pre-training, visual pretraining and multimodal pre-training, which requires more computing resources. Besides, FITE requires pre-generated facial expression descriptions and is not suitable for the sample

Method	Twitter-2015		Twitter-2017	
	Acc	Macro-F1	Acc	Macro-F1
A^2II	79.46	75.16	74.39	72.35
w/o Aspect	77.72	73.82	73.01	71.55
w/o Fusion	78.30	74.04	72.37	70.36
w/o Selector	77.82	72.55	71.72	70.24

Table 3: Ablation results on Twitter-2015 and Twitter-2017 dataset.

without facial emotion in the visual modality. Furthermore, most of the above methods are aimed at discriminative classification tasks which make them difficult to expand to other tasks, such as multi-modal aspect extraction task and multi-modal aspect-sentiment pair extraction task. However, our adopted model only has 437M parameters (not including the frozen vision encoder) which does not consume too many computing resources, and it is also an end-to-end generative model, which is easier to extend to other tasks.

4.4. Ablation Study

In this section, we conduct ablation studies to analyse the effect of the individual components of our method on Twitter-2015 and Twitter-2017 datasets. We report results of the following settings: directly using the entire twitter contents instead of using aspects to fuse with image features (wo-Aspect), reinitialize parameters of Q-Former without using the fusion ability of large language-vision model (wo-Fusion), remove the selector module (wo-Selector). The results are shown in Table 3.

4.4.1. Effect of Aspect

According to Table 3, by directly using the entire twitter contents to Q-Former, the accuracy and F1-score in Twitter-2015 dropped by 1.74 point and 1.34 point respectively. The accuracy and F1-score in Twitter-2017 also dropped by 1.38 point and 0.80 point respectively. We speculate that this is because inputting aspects can make the model pay more attention to the fine-grained features related to aspects in the image, and inputting the whole twitter sentences may introduce redundant noise information in the image which is related to sentences but not related to aspects. This observation indicates that it is necessary to use fine-grained aspect information in the fusion of images and texts.

4.4.2. Effect of Fusion

To verify the strength of fusion ability of large language-vision model, we also investigate the performance of the model when the Q-former model parameters are reinitialized and then fine-tuned it using only the current domain multimodal data. The experimental results show that the accuracy and F1-score in Twitter-2015 decreased by 1.16 points and 1.12 points respectively, and the accuracy and F1-score in Twitter-2017 decreased by 2.02 points and 1.99 points respectively.This demonstrate that only using current domain data which is relatively small to make the model learn cross-modality ability is limited, and the large language-vision model learns through large-scale data in multiple domains, which makes its ability to fuse images and text stronger. The Q-Former module of instructBLIP (Dai et al., 2023) provides a good bridge for image and text fusion while the number of parameters is relatively small.

4.4.3. Effect of Selector

We also explore the influence of selector module, and the model significantly performs worse without it. When the selector module is removed, the model cannot autonomously select the appropriate instruction for the language model. This means the same instruction will be used in all cases, causing the model to make judgments based on image information even when the image and text are irrelevant. This prevents the model from filtering out image noise through instructions. We can find that removing the selector module leads to a decline of 1.64 points and 2.61 points in accuracy and F1 score in Twitter-2015, and a decline of 2.67 points and 2.11 points in the accuracy and F1 score in Twitter-2017. This validates that selector module helps to reduce irrelevant image noise.

4.5. Further Analysis

To further demonstrate the effectiveness of our method, we conduct a comprehensive analysis in different scenarios. To be specific, we further tested the model in the case of completely image noise scenario, few-shot scenario, and simply using fine tuning without instructions.

4.5.1. Analysis of Image Noise

In order to study the influence of image noise and the performance of different models when images and aspects are unrelated, we randomly shuffled all the images while keeping the tweets and aspects unchanged, which is equivalent to converting the original dataset of partially image-text unrelated scenarios to all image-text unrelated scenarios, we then test CapBert, FITE and our model in the shuffled dataset. The experimental results are shown in Figure 2. It can be found that the Macro-F1 score of CapBert and FITE have decreased significantly in the scene where the image and text are totally unrelated, while our model has only a slight

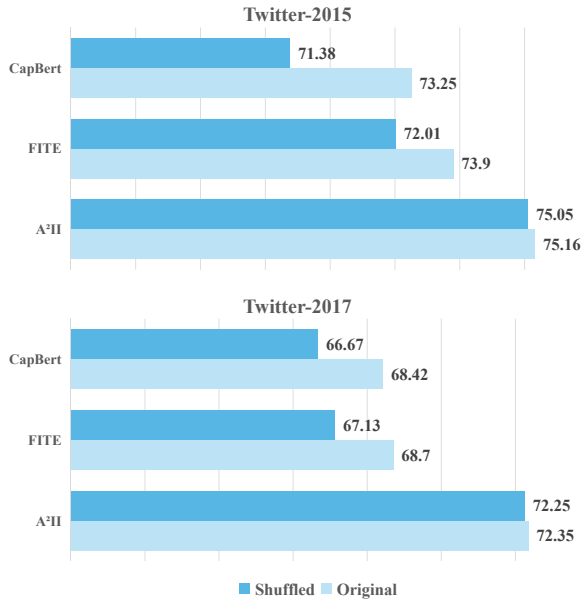

Figure 2: Comparison of Marco-F1 scores on shuffled datasets.

decrease. Based on the results, we can make a couple of observations: (1)The performance of the three models has declined, indicating that image noise has a negative impact on the final sentiment predictions to a certain extent. But the performance does not drop too much, which indicates that the text modality still plays a major role for multimodal sentiment classification. (2)The decline of CapBert and FITE is more significantly and it may be due to the fact that CapBert mainly uses image captions to help language models recognize aspect sentiments, while FITE additionally uses face descriptions, which makes the model more dependent on related images. However, when images are not related to the texts, this strong dependence will reduce the robustness of the model. (3)The performance of our model drops the least, which may be due to the fact that the selector tends to choose instructions that make the model not pay attention to the image information when the image and text are not relevant, thereby reducing the impact of image noise on the final predictions to a certain extent.

4.5.2. Analysis of Few Shot

We also further test the performance of our model in few-shot scenarios. Since labeling a large amount of multimodal fine-grained data requires a lot of manpower and resources, some research work focusing on multimodal sentiment classification in few-shot scenarios has recently emerged. PVLM (Yu and Zhang, 2022) proposes a prompt-based vision-aware language modeling approach to MABSC in the multimodal few-shot setting, UP-

Dataset	Full Split			Few-Shot Split		
	Train	Dev	Test	Train	Dev	Test
Twitter-2015	3179	1122	1037	36	36	1037
Twitter-2017	3562	1176	1234	36	36	1234

Table 4: Basic statistics of few-shot datasets for multimodal aspect-oriented sentiment classification task.

Method	Twitter-2015		Twitter-2017	
	Acc	Macro-F1	Acc	Macro-F1
PVLM	60.85	47.65	52.29	51.14
UP-MPF	56.03	53.06	52.61	52.44
A^2II (Ours)	**63.16**	**61.24**	**53.97**	**53.30**

Table 5: Experiment results for few-shot multimodal aspect-oriented sentiment classification task.

MPF (Yu et al., 2022b) introduces a unified pre-training for multimodal prompt-based fine-tuning, which is a strong baseline for few-shot MABSC. We also use the dataset like Yu and Zhang (2022) and Yu et al. (2022b), keep the development set and test set unchanged, then randomly select 1% from the training set according to the label distribution as the labeled data in the few-shot setting. The statistical information of the few-shot datasets is shown in Table 4.

According to experiments in few-shot setting, we find that our model outperforms two baseline methods which are specifically introduced for fewshot scenarios. On Twitter15, the accuracy has 2.31 point improvement (60.85→63.16) and the macro F1-score has 8.18 point improvement (53.06→61.24) compared with the previous best methods PVLM and UP-MPF. On Twitter17, the accuracy has 1.36 point improvement (52.61→53.97) and the macro F1-score has 0.86 point improvement (52.44→53.30) compared with the previous best method UP-MPF. We speculate that this effect might result from the utilization of instruction tuning, which enhances the model's transferability. When the number of annotated training samples is relatively small, the model can also exploit them to learn through instruction tuning and use its pre-trained internal knowledge to make correct predictions, which further emphasizes that instruction tuning plays an important role in MABSC tasks.

4.5.3. Analysis of Instruction Tuning

In order to further explore the role of instruction tuning paradigm and to illustrate that the performance improvement is due to the effect of instruction tuning, we also fine-tuned a model without using instructions and compared the results with our

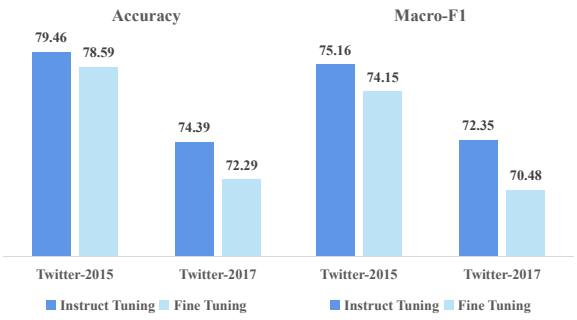

Figure 3: Comparison of fine-tuning and instruction tuning for MABSC task.

Label	Negative	Positive
Image		
Sentence	"Meek Mill beats Drake, Kendrick Lamar , Future to win Top Rap Album at Billboard Music Awards"	RT @ CBSSunday : Gallery : The haunting beauty of winter in Yellowstone.
Instruction	Definition: Based solely on the information in the following sentence to identify the sentiment of aspect in the sentence.	Definition: Combining information from image and the following sentence to identify the sentiment of aspect in the sentence.
Aspect	Kendrick Lamar	Yellowstone
CapBert	Positive (x)	Neutral (x)
FITE	Positive (x)	Neutral (x)
A²II (ours)	Negative (√)	Positive (√)

Figure 4: Predictions of CapBERT, FITE, and our model on several test samples.

model using instruction tuning. The experimental results are shown in Figure 3. Experiments show that when we add instruction prompts specific to the downstream MABSC task in the form of task definition, the model achieves higher accuracy and macro-F1 score than models only using vanilla fine-tuning, which indicates that using instruction tuning is more effective than just using vanilla fine-tuning in MABSC task. And we hope that this observation will encourage further research in this direction.

5. Case Study

To demonstrate the influence of image noise and better understand the advantage of our model, figure 4 shows the comparison between the predictions of the CapBERT, FITE and our model on two representative test samples. First, on the left side of Figure 4, given the aspect Kendrick Lamar, our model accurately predicted its sentiment as Negative, while CapBERT and FITE incorrectly predicted its sentiment as Positive. We speculate that this may be due to the fact that CapBERT and FITE translated the unrelated image to captions and Meek Mill's facial descriptions which are not related to the aspect Kendrick Lamar, thereby affecting the prediction of the model. However, our model gave the correct sentiment prediction by instructing the model to ignore unrelated image information and based solely on the sentence information to identify the sentiment of the aspect. In addition, on the right side of Figure 4, given the aspect Yellowstone, our model accurately predicted its sentiment as Positive, while CapBERT and FITE incorrectly predicted its sentiment as Neutral. This may be because the image lacks facial emotion, making it unsuitable for FITE, furthermore, CapBERT is more inclined to generate captions with neutral sentiment, while our model fuses images and aspects better. The selector module selected the appropriate instruction to combine image information and aspect, resulting in a positive sentiment prediction for the aspect.

6. Conclusion

In this paper, we explore an instruction tuning modeling approach for multimodal aspect-oriented sentiment classification task. To tackle the challenge of potentially irrelevant images and aspects, we propose a selector module to mitigate the impact of image noise. To alleviate the limitation of cross-modal fusion, we leverage the ability of large language-vision models without introducing too many additional parameters. Experiment results on two MABSC datasets and two few-shot datasets show that our method outperforms a series of benchmark models and demonstrate the effectiveness of our method for achieving crossmodal alignment within multimodal data and its robustness in image-aspect irrelevant scenario.

7. Acknowledgements

We thank the anonymous reviewers for their constructive comments and suggestions.

8. Bibliographical References

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022a. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022b. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Junnan Li, Dongxu Li, AnthonyMeng Huat, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500v2*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2149–2159.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *Learning,Learning*.

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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 conference on empirical methods in natural language processing*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019a. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. 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)*, pages 380–385.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019b. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL-HLT*, pages 380–385.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224.

Jiawei Wang, Zhe Liu, Victor Sheng, Yuqing Song, and Chenjian Qiu. 2021. Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4*, pages 3–15. Springer.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016a. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016b. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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

Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3324–3335.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. IJCAI.

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

Jianfei Yu, Jieming Wang, Rui Xia, and Junjie Li. 2022a. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *Proc. of the Thirty-First Int. Joint Conf. on Artificial Intelligence, IJCAI 2022*, pages 4482–4488.

Yang Yu and Dong Zhang. 2022. Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Yang Yu, Dong Zhang, and Shoushan Li. 2022b. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 189–198.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.