DEQA: Descriptions Enhanced Question-Answering Framework for Multimodal Aspect-Based Sentiment Analysis

Zhixin Han[1], Mengting Hu[1]*, Yinhao Bai[2], Xunzhi Wang[1], Bitong Luo[1]
[1]College of Software, Nankai University
[2]JD AI Research, Beijing, China
zhixinhan@mail.nankai.edu.cn, mthu@nankai.edu.cn, yinhaobai68@gmail.com,
xunzhi@mail.nankai.edu.cn, luobitong@mail.nankai.edu.cn

Abstract

Multimodal aspect-based sentiment analysis (MABSA) integrates text and images to perform fine-grained sentiment analysis on specific aspects, enhancing the understanding of user opinions in various applications. Existing methods use modality alignment for information interaction and fusion between images and text, but an inherent gap between these two modalities necessitates a more direct bridging mechanism to effectively connect image understanding with text content. For this, we propose the Descriptions Enhanced Question-Answering Framework (DEQA), which generates descriptions of images using GPT-4, leveraging the multimodal large language model to provide more direct semantic context of images. In DEQA, to help the model better understand the task's purpose, we frame MABSA as a multi-turn question-answering problem to add semantic guidance and hints. We input text, image, and description into separate experts in various combinations, allowing each expert to focus on different features and thereby improving the comprehensive utilization of input information. By integrating these expert outputs within a multi-turn question-answering format, we employ a multi-expert ensemble decision-making approach to produce the final prediction results. Experimental results on two widely-used datasets demonstrate that our method achieves state-of-the-art performance. Furthermore, our framework substantially outperforms GPT-4o and other multimodal large language models, showcasing its superior effectiveness in multimodal sentiment analysis.

Introduction

Multimodal aspect-based sentiment analysis (MABSA) (Ju et al. 2021) is an advanced field at the intersection of natural language processing and computer vision, integrating text and images to perform fine-grained sentiment analysis on specific aspects, enhancing the understanding of user opinions in various applications. As shown in Figure 1, MABSA involves extracting all aspect terms from image-text pairs and predicting their sentiment polarities. It includes two subtasks: multimodal aspect term extraction (MATE) (Wu et al. 2020a), which identifies all aspect terms in the sentence prompted by the associated image; and multimodal aspect sentiment classification (MASC) (Yu and Jiang 2019),

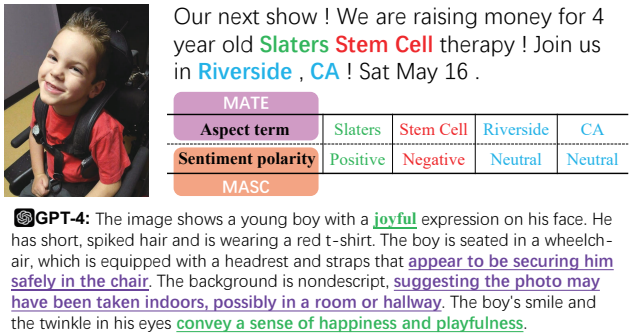

Figure 1: A dataset entry from Twitter2015 (Yu and Jiang 2019). MABSA contains two subtasks, MATE and MASC.

which determines the sentiment polarity of each aspect. MABSA requires the utilization of both image and text information, take Figure 1 as an example, where solely based on the text, people might assume the sentiment polarity of *Slaters* to be *negative* because he is ill and requires treatment. However, upon observing the image, it becomes clear that the young boy is smiling. In this case, we can infer that the sentiment polarity of *Slaters* is *positive*.

Image and text, as two distinct modalities, exhibit significant differences in the types of information they convey and the forms in which they are expressed, each carrying heterogeneous information. Some methods focus on fusing these two modalities, while others advocate for modality alignment as the main means to effectively integrate the heterogeneous information, thereby improving performance. Ling, Yu, and Xia (2022) introduce three task-specific pre-training tasks to identify fine-grained aspect, opinions, and their cross-modal alignments. Zhou et al. (2023) introduce aspect-aware attention module (A^3M) for semantically fine-grained image-text alignment. Furthermore, to achieve better alignment, many cutting-edge methodologies adopt a pre-training followed by fine-tuning approach. However, despite the advancements in modality alignment, the inherent gap between image and text modalities still exists, necessitating a more direct bridging mechanism to effectively connect image understanding with text content. What is more, the pre-training process consumes a significant amount of time and computational resources.

*Corresponding author.

For this, we propose the **D**escriptions **E**nhanced **Q**uestion-**A**nswering Framework (DEQA), which generates descriptions of images using GPT-4, leveraging the multimodal large language model to provide more direct semantic context of images. Unlike previous methods, our method does not require pre-training, significantly reducing the demand for computational resources. Furthermore, we find that the descriptions not only contain purely visual information but also incorporate related knowledge, commonsense, and appropriate inferences made by GPT-4 based on the image content. As illustrated in Figure 1, in the description, the underlined portions are supplementary knowledge and inferences made by GPT-4, and some of them can aid our model in making predictions. Terms *joyful* and *convey a sense of happiness and playfulness* suggest that the sentiment polarity of the aspect term *Slaters* is likely to be *positive*.

We design two sub-models to separately handle the MATE and MASC tasks, and then sequentially connect the inputs and outputs of these two sub-models to complete the MABSA task. Each sub-model contains three experts, with each dedicated to processing one of the following scenarios: text-only input, text and image input, and text and descriptions input, allowing each expert to specialize in different features. This allows us to comprehensively utilize these three types of information. Furthermore, to help the model better understand the task's purpose and context, we frame MABSA as a multi-turn question-answering problem to add semantic guidance and hints. By integrating expert outputs within this question-answering format, we employ multi-expert ensemble decision-making approaches to produce the final prediction results. Experimental results on two widely-used datasets demonstrate that our method achieves state-of-the-art performance. What is more, our framework substantially outperforms GPT-4o and other multimodal large language models, showcasing its superior effectiveness in multimodal sentiment analysis.

In summary, our contributions are as follows:

- We introduce the DEQA for MABSA. This framework leverages the capabilities of GPT-4 to generate descriptions from images. DEQA integrates queries, transforming MABSA into the structured question-answering format. We input text, image, and description into separate experts in various combinations, allowing each expert to focus on different features. By integrating these expert outputs within the multi-turn question-answering format, we employ multi-expert ensemble decision-making approaches to produce the final prediction results.
- Our method demonstrates state-of-the-art performance on two widely-used datasets, namely, Twitter2015 and Twitter2017 (Yu and Jiang 2019).
- We evaluate `gpt-4o-2024-05-13` on the MABSA task, and compare our model with it and other commonly used multimodal large language models. The results demonstrate that DEQA significantly outperforms GPT-4o as well as other multimodal large language models.

Related Work

Given the prevalence of multimodal data on the Internet, information from the visual modality can be utilized to provide complementary sentiment signals to text features (Zhang, Wang, and Liu 2018). Thus, MABSA and its subtasks are widely studied. Next, we will introduce some classic and cutting-edge methods that also serve as the baselines for comparison.

Aspect-Based Sentiment Analysis

Unlike MABSA, aspect-based sentiment analysis (Pontiki et al. 2014) relies solely on the text modality. Existing methods primarily focus on capturing the structural information within the text, aiming to extract richer semantic details from its structure and relationships. Hu et al. (2019) propose a span-based extract-then-classify framework and denote it as SPAN, Chen, Tian, and Song (2020) propose directional graph convolutional networks (D-GCN), and Yan et al. (2021) exploit the pre-training sequence-to-sequence model BART (Lewis et al. 2019) to solve all aspect-based sentiment analysis subtasks in an end-to-end framework. We compare these three text-based methods with our multimodal approach on MABSA to highlight the advantages of "multimodal" and our method.

Multimodal Aspect Term Extraction

MATE aims to extract all aspect terms from the given text-image pair. To address this task, Wu et al. (2020b) propose a region-aware alignment network (RAN). OCSGA (Wu et al. 2020c) and UMT (Yu et al. 2020) are originally proposed for multimodal named entity recognition (Moon, Neves, and Carvalho 2018). However, due to the high degree of similarity between multimodal named entity recognition and the MATE task, these methods can also be effectively adapted for MATE.

Multimodal Aspect Sentiment Classification

Given an aspect term, MASC aims to identify the corresponding sentiment polarity from the text-image pair. Given that the aspect term is already known, some existing methods focus on aligning the specific aspect term with the image to selectively extract the image region information relevant to the aspect term. Yu, Jiang, and Xia (2020) propose an entity-sensitive attention and fusion network (ESAFN), Yu and Jiang (2019) propose the target-oriented multimodal BERT (TomBERT), and Khan and Fu (2021) introduce a two-stream model that translates images in input space and leverages this translation to construct an auxiliary sentence that provides multimodal information.

Multimodal Aspect-based Sentiment Analysis

MABSA aims to extract all aspect terms from the image-text pair and predict their sentiment polarities. Ju et al. (2021) carries out the MABSA task by combining, transferring, and modifying existing models that are not originally designed for MABSA. Specifically, Ju et al. (2021) implement two pipeline approaches upon two representative studies of MATE and MASC and three collapsed tagging ap-

proaches. 1) UMT+TomBERT. 2) OCSGA+TomBERT. 3) UMT-collapsed. 4) OCSGA-collapsed. 5) RpBERT.

The following five models are specifically designed for MABSA and all its subtasks. Ju et al. (2021) are the first to jointly perform MATE and MASC, proposing a multi-modal joint learning approach, namely JML. Ling, Yu, and Xia (2022) propose a task-specific vision-language pre-training framework for MABSA (VLP-MABSA). Yang, Na, and Yu (2022) propose a multi-task learning framework named cross-modal multitask transformer (CMMT). Zhou et al. (2023) propose an aspect-oriented method (AoM) to detect aspect-relevant semantic and sentiment information. Peng et al. (2024) propose a framework called DQPSA, which contains a prompt as dual query module and an energy-based pairwise expert module.

Methodology

Framework Overview

Given a tweet that contains an image I and a sentence S, DEQA aims to identify the set of pairs $\{(a_1, s_1), (a_2, s_2), \ldots, (a_i, s_i), \ldots\}$. Here, (a_i, s_i) represents a pair of (aspect term, sentiment polarity), and s_i belongs to the set $\{positive, neutral, negative\}$.

As shown in Figure 2, to formalize MABSA as multiple-instance question-answering tasks, we construct three types of queries for a pair, including an aspect extraction query Q^e, an aspect validation query Q^v, and a sentiment classification query Q^c. Concretely, at first, the aspect extraction query Q^e aims to extract the aspect term a_i from the sentence S. Then, given the aspect term a_i, the aspect validation query Q^v is designed to validate the accuracy of a_i. Finally, the sentiment classification query Q^c aims to predict the sentiment polarity s_i for a_i. Additionally, by expressing questions in natural language, semantic hints are provided to the model, aiding in a better understanding of the tasks' objectives.

It is worth noting that, in Figure 2, we introduce two special tokens, <target> and </target>, to mark a_i, emphasizing and differentiating this aspect term. We find that some sentences contain multiple instances of the same aspect term, and marking them helps to avoid ambiguity. For instance, in the Twitter2015 dataset, there is a sentence, *RT @SoSingaporean: What people from other countries do at IKEA VS What I do at IKEA #sosingaporean*, where *IKEA* appears twice.

Our model contains two sub-models, one for MATE, and the other for MASC. In each sub-model, there are three experts: the text-only expert, the text and description expert, and the text and vision expert, each responsible for processing different combinations of modality inputs. Additionally, each sub-model includes a decision ensemble that integrates the outputs from the various experts to determine the final prediction. Finally, the output of the sub-model for MATE serves as the input for the sub-model for MASC, and by combining the outputs of these two sub-models, the final prediction for the MABSA task is obtained.

Sub-model for MATE

The sub-model starts with the text-only aspect extraction expert, which utilizes the aspect extraction query Q^e to extract the aspect term a_i using a pre-trained language model combined with the BIO tagging scheme (Huang, Xu, and Yu 2015) and CRF (Lafferty et al. 2001). Following this, the text and description aspect validation expert incorporates both text and description to validate the extracted aspect terms with the help of the aspect validation query Q^v. Meanwhile, the text and vision aspect validation expert utilizes Q^v and integrates visual information with text to confirm the correctness of aspect terms. Finally, the aspect extraction decision ensemble combines predictions from two validation experts, filtering the extracted aspect terms to determine final predictions.

Text-Only Aspect Extraction Expert We construct the input as shown in Figure 2, and feed it into DeBERTa (He et al. 2021) to obtain a representation for each token. These token representations are then passed through a fully connected layer to produce $\mathbf{X}$, which is used for predicting the labels according to the BIO tagging scheme. Subsequently, we use CRF to compute the conditional probability of the label sequence $\mathbf{y}$ (the gold labels) given $\mathbf{X}$:

$$P(\mathbf{y} \mid \mathbf{X}) = \frac{\exp(\text{score}(\mathbf{X}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}} \exp(\text{score}(\mathbf{X}, \mathbf{y}'))} \quad (1)$$

where $\mathcal{Y}$ represents the set of all possible label sequences, and score $(\cdot)$ is the score function (Lafferty et al. 2001) that assigns a score to the label sequence $\mathbf{y}$ based on $\mathbf{X}$. The loss function is the negative log-likelihood of this conditional probability:

$$\mathcal{L}_{\text{t}}^{a} = -\log P(\mathbf{y} \mid \mathbf{X}) \quad (2)$$

Text and Description Aspect Validation Expert We construct the input as shown in Figure 2, and feed it into DeBERTa to obtain the representation for each token. To ensure that the model effectively utilizes the functionality and role of the added special tokens, we deviate from the common practice of selecting the <s> token. Instead, we select the representation of the first <target> token and pass it through a fully connected layer to predict whether the aspect term is correct. The output of the fully connected layer is then applied to a softmax function (Bridle 1989) to obtain the predicted probabilities $\mathbf{P}_{\text{d}}^{a}$. Finally, we compute the cross-entropy loss (Rumelhart, Hinton, and Williams 1986):

$$\mathcal{L}_{\text{d}}^{a} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (3)$$

where N is the number of classes (i.e., correct or incorrect), y_i is the true label for class i, and $\hat{y}_i$ is the predicted probability for class i.

Text and Vision Aspect Validation Expert As shown in Figure 2, we feed the aspect validation query Q^v into the CLIP (Radford et al. 2021) text encoder to obtain the representation $\mathbf{t}$ of the entire text. We feed the image I into the CLIP vision encoder to obtain the representation

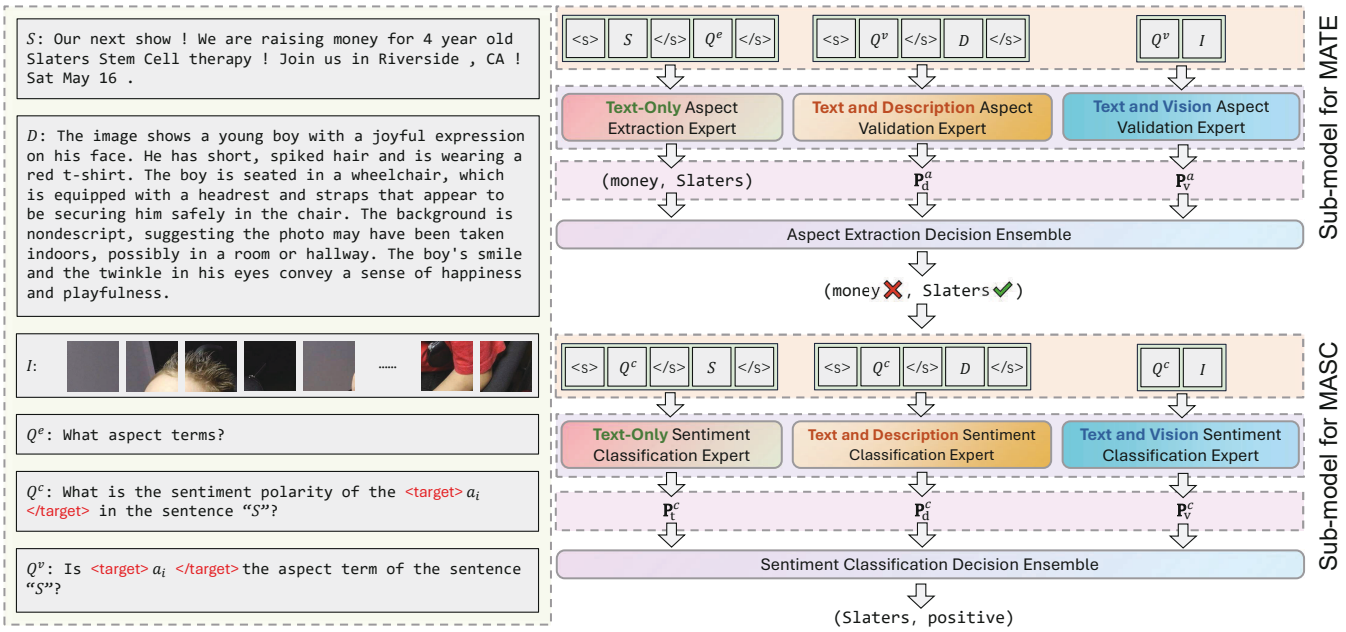

Figure 2: Framework of DEQA. <s> is the beginning token, and </s> is the segment token (He et al. 2021). The determined aspect term in S is enclosed by two special tokens, <target> and </target>, except for the input of the text-only aspect extraction expert. For the text-only sentiment polarity expert, Q^c is *What is the sentiment polarity of the* <target> *Slaters* </target>*?*.

$\mathbf{V} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n]$ for patch sequence. We employ an attention mechanism (Vaswani et al. 2017) to achieve cross-modal alignment between the text and image patches, using $\mathbf{t}$ as the query and $\mathbf{p}_i$ as the key and value. This approach ultimately yields an aligned representation of the question and the image:

$$\mathbf{g} = \mathrm{CA}(\mathbf{t}, \mathbf{V}) \tag{4}$$

where $\mathrm{CA}(\cdot)$ denotes multi-head cross attention (Chen, Fan, and Panda 2021). Unlike conventional methods, we utilize multimodal factorized bilinear pooling (MFB) (Yu et al. 2017) as the attention scoring function. Subsequently, we convert the original text representation:

$$\mathbf{t}' = \mathrm{ReLU}(\mathbf{W} \cdot \mathrm{Dropout}(\mathbf{t}, 0.5) + \mathbf{b}) \tag{5}$$

where $\mathbf{W}$ is the weight matrix, $\mathbf{b}$ is the bias vector, and $\mathrm{Dropout}(\mathbf{t}, 0.5)$ applies dropout (Srivastava et al. 2014) with a probability of 0.5 to the original text representation. The function $\mathrm{ReLU}(\cdot)$ is the rectified linear unit activation function (Krizhevsky, Sutskever, and Hinton 2012). We then employ MFB again to fuse $\mathbf{t}'$ and $\mathbf{g}$, obtaining the final fused representation:

$$\mathbf{f} = \mathrm{MFB}(\mathbf{t}', \mathbf{g}) \tag{6}$$

We pass $\mathbf{f}$ through a fully connected layer to predict whether the corresponding aspect term is correct. The output of the fully connected layer is then passed through a softmax function to obtain the predicted probabilities $\mathbf{P}^a_{\mathrm{v}}$. Finally, we derive the cross-entropy loss $\mathcal{L}^a_{\mathrm{v}}$.

Aspect Extraction Decision Ensemble We combine $\mathbf{P}^a_{\mathrm{d}}$ with $\mathbf{P}^a_{\mathrm{v}}$, by element-wise addition and normalization to obtain the final probability distribution for each label:

$$\mathbf{P}^a = \frac{\mathbf{P}^a_{\mathrm{d}} + \mathbf{P}^a_{\mathrm{v}}}{\|\mathbf{P}^a_{\mathrm{d}} + \mathbf{P}^a_{\mathrm{v}}\|_1} \tag{7}$$

Then, we use $\mathbf{P}^a$ to validate the extracted aspect terms.

To jointly train each expert in MATE sub-model and make them mutually beneficial, we sum the loss functions of the different experts to form the overall loss objective of the MATE sub-model:

$$\mathcal{L}^a = \mathcal{L}^a_{\mathrm{t}} + \mathcal{L}^a_{\mathrm{d}} + \mathcal{L}^a_{\mathrm{v}} \tag{8}$$

Sub-model for MASC

This sub-model predicts the sentiment polarity for each aspect term a_i determined by the sub-model for MATE. The text-only sentiment classification expert uses a pre-trained language model to predict the sentiment polarity. Meanwhile, the text and description sentiment classification expert refines sentiment polarity predictions by incorporating text and description, and the text and vision sentiment classification expert integrates visual data with text to make predictions. Finally, the sentiment classification decision ensemble aggregates outputs from the three experts to determine the final sentiment polarity for each aspect term.

Text-Only Sentiment Classification Expert We construct the input as shown in Figure 2, and feed it into DeBERTa to obtain the representation for each token. Then, we use the first <target> token to predict the sentiment polarity, and obtain the predicted probabilities $\mathbf{P}^c_{\mathrm{t}}$. Finally, we derive the cross-entropy loss $\mathcal{L}^c_{\mathrm{t}}$.

Methods	Twitter2015			Twitter2017		
	P	R	F_1	P	R	F_1
RAN†	80.5	81.5	81.0	90.7	90.0	90.3
UMT†	77.8	81.7	79.7	86.7	86.8	86.7
OCSGA†	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
DQPSA	**88.3**	87.1	87.7	**95.1**	93.5	94.3
DEQA	86.6	**89.5**	**88.0**	93.8	**95.1**	**94.4**

Table 1: Results of different methods for MATE. F_1 denotes Micro-F1. † denotes the results from (Ju et al. 2021). The best results are bold-typed and the second best ones are underlined.

Text and Description Sentiment Classification Expert We construct the input as shown in Figure 2, and feed it into DeBERTa to obtain the representation for each token. Then, we use the first `<target>` token to predict the sentiment polarity, and obtain the predicted probabilities $\mathbf{P}_{\mathrm{d}}^{c}$. Finally, we derive the cross-entropy loss $\mathcal{L}_{\mathrm{d}}^{c}$.

Text and Vision Sentiment Classification Expert As shown in Figure 2, we feed the sentiment classification query Q^c and the image I into CLIP. Subsequently, we adopt the same approach as used in the text and vision aspect validation expert to obtain $\mathbf{P}_{\mathrm{v}}^{c}$. Finally, we derive the cross-entropy loss $\mathcal{L}_{\mathrm{v}}^{c}$.

Sentiment Classification Decision Ensemble Similar to the aspect extraction decision ensemble, we obtain the final probability distribution:

$$\mathbf{P}^{c}=\frac{\mathbf{P}_{\mathrm{d}}^{c}+\mathbf{P}_{\mathrm{v}}^{c}+\mathbf{P}_{\mathrm{t}}^{c}}{\|\mathbf{P}_{\mathrm{d}}^{c}+\mathbf{P}_{\mathrm{v}}^{c}+\mathbf{P}_{\mathrm{t}}^{c}\|_{1}} \tag{9}$$

Unlike the training approach used for the sub-model for MATE, we have trained each MASC expert individually. Specifically, we employ three separate losses—$\mathcal{L}_{\mathrm{t}}^{c}$, $\mathcal{L}_{\mathrm{d}}^{c}$ and $\mathcal{L}_{\mathrm{v}}^{c}$—to guide the training. In the sub-model for MATE, three experts have a strong sequential and logical relationship among them, which indicates significant mutual influence. Therefore, joint training can effectively account for the interactions among these experts. Conversely, in the sub-model for MASC, the relationships among the three experts are relatively independent. Thus, training each expert separately and performing individual hyperparameter tuning for each one are more beneficial.

Experiments

Datasets

Following previous studies, we use two widely adopted benchmarks: Twitter2015 and Twitter2017[1] to evaluate DEQA.

[1]The basic statistics of these two datasets are provided in Yu and Jiang (2019)'s paper.

Implementation Details

We use `gpt-4-vision-preview` to generate descriptions of the images. Our code and data[2] provide implementation details, including pre-trained models, training details, and training durations across two datasets.

We train our model using an NVIDIA RTX A6000 GPU and implement an early stopping strategy (Prechelt 1998) with a patience of 3 epochs and a threshold of 0.01 to prevent overfitting. The AdamW optimizer (Loshchilov and Hutter 2019) is utilized for training, with a weight decay (Krogh and Hertz 1991) of 0.01. Additionally, we employ a linear learning rate scheduler with a warmup ratio (He et al. 2015) of 0.1 to adjust the learning rate throughout the training process.

In the MASC sub-model, we first train the text-only sentiment classification expert. Subsequently, the fine-tuned weights of this expert are used to initialize the text and description sentiment classification expert. For the two text and vision experts, the factor dimension of MFB is set to 1 for attention scoring and 8 for fusion. During the training process, we freeze the CLIP image encoder.

Evaluation Metrics

In continuation of prior studies, we assess the performance of our model on the MATE and MABSA tasks using Precision (P), Recall (R), and Micro-F1 (F_1) scores. For the MASC task, we report both Accuracy (Acc) and Macro-F1 (F_1) scores.

Results

Performance on MATE Table 1 presents the comparative results for MATE. Our method surpasses the second-best model on both Twitter2015 and Twitter2017 datasets. Although our method offers only a slight advantage over DQPSA, it is important to note that DQPSA relies on computationally intensive pre-training.

Performance on MASC Table 2 presents the comparative results for MASC. Our method surpasses the second-best models on Twitter2015. However, the performance of DEQA on the Twitter2017 dataset is not particularly outstanding. In terms of F1 score, our method exceeds AoM by only 0.1%; meanwhile, in terms of accuracy, it falls behind AoM by 0.6%. Peng et al. (2024) point out that Twitter2017 contains a significant number of unresolvable and unidentifiable symbols, including emojis commonly used on Twitter, which are unknown to the DeBERTa model we used. Given this, we believe our proposed method remains effective for MASC.

Performance on MABSA Table 3 presents the comparative results for MABSA. Our method surpasses the second-best model on both Twitter2015 and Twitter2017 datasets. This demonstrates that our model has successfully achieved state-of-the-art performance. Furthermore, compared to text-based models, DEQA exhibits significantly better performance. Additionally, we construct an

[2]https://github.com/ZhixinHan/DEQA

Methods	Twitter2015		Twitter2017	
	Acc	F_1	*Acc*	F_1
TomBERT	77.2	71.8	70.5	68.0
ESAFN	73.4	67.4	67.8	64.2
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	<u>75.9</u>	**76.4**	<u>75.0</u>
DQPSA	<u>81.1</u>	81.1†	75.0	75.0†
DEQA	**82.1**	**77.6**	<u>75.8</u>	**75.1**

Table 2: Results of different methods for MASC. F_1 denotes Macro-F1. † denotes the use of Micro-F1 as the evaluation metric. The best results are bold-typed and the second best ones are underlined.

Methods	Twitter2015			Twitter2017		
	P	R	F_1	P	R	F_1
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*
UMT+TomBERT†	58.4	61.3	59.8	62.3	62.4	62.4
OCSGA+TomBERT†	61.7	63.4	62.5	63.4	64.0	63.7
OCSGA-collapse†	63.1	63.7	63.2	63.5	63.5	63.5
UMT-collapse†	60.4	61.6	61.0	60.0	61.7	60.8
RpBERT†	49.3	46.9	48.0	57.0	55.4	56.2
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	<u>71.0</u>	69.7
DQPSA	**71.7**	<u>72.0</u>	<u>71.9</u>	<u>71.1</u>	70.2	<u>70.6</u>
DEQA (end-to-end)	64.0	64.7	64.4	64.3	63.9	64.1
DEQA	<u>71.4</u>	**73.9**	**72.7**	**71.4**	**72.4**	**71.9**

Table 3: Results of different methods for the MABSA task. F_1 denotes Micro-F1. † denotes the results from (Ju et al. 2021); ‡ denotes the results from (Ling, Yu, and Xia 2022). The best results are bold-typed and the second best ones are underlined; the best results for text-based methods are italicized.

end-to-end version of DEQA for comparison. Specifically, we use `deberta-v3-large` for the text modality and `clip-vit-large-patch14-336` for the image modality, with cross-attention as the fusion method.

Ablation Study

W/o Aspect Validation Query After removing aspect validation query Q^v, as shown in Table 4, a decline in the F1 score is observed for the MATE task across both datasets. Specifically, for the Twitter2015 dataset, precision decreases by 1.4%, while recall increases by 0.2%. The increase in recall indicates more correct aspect terms being identified, specifically an increase in true positives and a decrease in false negatives. However, despite the increase in true positives, the precision shows a larger decrease, suggesting an increase in false positives. Based on the above reasoning, we can conclude that, **on the one hand**, Q^v may lead to fewer correct aspect terms being identified, indicating that Q^v might misclassify some correctly identified aspect terms as incorrect. **On the other hand**, Q^v can result in a reduction in false positives, implying that Q^v effectively filters out some incorrect predictions. Overall, when removing Q^v, the larger decrease in precision compared to the smaller increase in recall suggests that the misclassification of correct predictions by Q^v is less significant than its role in filtering out incorrect predictions. For the Twitter2017 dataset, both precision and recall decrease, but the drop in precision (1.5%) is greater than the drop in recall (0.4%), showing a similar trend to that observed in the Twitter2015 dataset.

W/o Semantic Hints We remove semantic hints from all queries. Specifically, we completely remove Q^e. We transform Q^v into <target> a_i </target> *of the sentence "S"*, and Q^c into <target> a_i </target> *in the sentence "S"*[3]. As shown in Table 4, the model's performance on the MATE task experience a slight decline, while its performance on the MASC and MABSA tasks see a noticeable drop. This suggests that semantic hints are important for MABSA and MASC but have a relatively minor effect on the MATE task.

W/o Special Tokens We remove the special tokens <target> and </target> from S and replace those not in S with quotation marks. Additionally, following common practice, we select the representation of the beginning token <s>. As shown in Table 4, the experimental results indicate that the special tokens we introduced play an important role.

W/o Description After removing descriptions, we observe a decline in model performance across all three tasks, as shown in Table 4. This indicates that descriptions are beneficial to MABSA.

W/o Vision We remove images, remaining descriptions. The experimental results, as shown in Table 4, indicate that removing images leads to declines in model performance. This suggests that descriptions cannot fully substitute for the images. While descriptions provide accurate and comprehensive summaries of the image content, they still fail to capture all the information that images offer, particularly detailed information. We observe that the performance degradation caused by the absence of descriptions is generally greater than that caused by the absence of visual modality. This indicates that descriptions are more important than images.

Sentiment Classification Decision Ensemble We evaluate several alternative decision methods. However, none outperforms the Sentiment Classification Decision Ensemble. Table 7 are results comparing different strategies on MASC.

[3]For the text-only sentiment classification expert, the query is transformed into <target> a_i </target>.

Tasks	Methods	Twitter2015				Twitter2017			
		Acc	*P*	*R*	F_1	*Acc*	*P*	*R*	F_1
MATE	**DEQA**	-	**86.6**	**89.5**	**88.0**	-	**93.8**	**95.1**	**94.4**
	w/o Aspect validation query	-	85.2	89.7	87.4	-	92.3	94.7	93.4
	w/o Semantic hints	-	86.9	88.2	87.6	-	94.2	94.7	94.4
	w/o Special tokens	-	86.9	87.7	87.3	-	93.2	95.1	94.1
	w/o Description	-	86.5	88.8	87.6	-	93.4	93.8	93.6
	w/o Vision	-	87.1	88.0	87.5	-	93.8	94.2	94.0
MASC	**DEQA**	**82.1**	-	-	**77.6**	**75.8**	-	-	**75.1**
	w/o Semantic hints	80.3	-	-	76.1	74.2	-	-	73.4
	w/o Special tokens	79.1	-	-	74.4	74.1	-	-	73.1
	w/o Description	79.8	-	-	74.4	72.2	-	-	71.0
	w/o Vision	80.9	-	-	77.0	75.7	-	-	74.7
MABSA	**DEQA**	-	**71.4**	**73.9**	**72.7**	-	**71.4**	**72.4**	**71.9**
	w/o Semantic hints	-	69.7	70.8	70.2	-	70.2	70.5	70.3
	w/o Special tokens	-	69.7	70.4	70.1	-	69.1	70.6	69.8
	w/o Description	-	68.9	70.8	69.9	-	68.0	68.3	68.2
	w/o Vision	-	71.0	71.7	71.3	-	71.1	71.5	71.3

Table 4: Results of ablation study. For the MATE and MABSA tasks, F_1 denotes Micro-F1, whereas for the MASC task, F_1 denotes Macro-F1.

Large Models	Twitter2015	Twitter2017
ChatGPT-3.5	<u>65.5</u>	60.0
LLaMA2-13B	60.4	48.5
Mixtral-AWQ	55.5	<u>60.2</u>
GPT-4V	*53.9*	*<u>60.2</u>*
Claude3-V	38.5	54.5
Gemini-V	54.5	59.3
LLaVA-v1.6-13B	58.7	56.1
Fuyu-8B	58.8	50.8
Qwen-VL-Chat	<u>65.5</u>	59.7
DEQA	**82.1**	**75.8**

Table 5: Results of different (multimodal) large language models for the MASC task. We use Accuracy as the evaluation metric. All results are from (Yang et al. 2024). The best results are bold-typed and the second best ones are underlined.

Models	Twitter2015			Twitter2017		
	P	*R*	F_1	*P*	*R*	F_1
gpt-4o-2024-05-13	23.7	28.5	25.9	25.2	26.3	25.7
DEQA	**71.4**	**73.9**	**72.7**	**71.4**	**72.4**	**71.9**

Table 6: Results of different models for MABSA. F_1 denotes Micro-F1.

Performance Compared to Large Models

We compare DEQA on the MASC task with several (multimodal) large language models. The results are presented in Table 5. It can be observed that our DEQA significantly outperforms all the (multimodal) large language models, including GPT-4V, LLaMA2-13B (Touvron et al. 2023), Mixtral-AWQ (Egiazarian et al. 2024) and Gemini-V (Qi et al. 2023). We also evaluate `gpt-4o-2024-05-13` on MABSA. The results are presented in Table 6. It can be observed that its performance is relatively poor compared to our method.

Methods	Twitter2015		Twitter2017	
	Acc	F_1	*Acc*	F_1
MF	79.6	74.3	74.1	73.3
MLR	81.2	76.0	74.6	73.5
PoE	81.2	76.7	75.6	74.9
MLP	81.9	77.4	75.0	74.3
PV	82.1	77.6	75.8	74.8
SCDE	**82.1**	**77.6**	**75.8**	**75.1**

Table 7: Results of different decision methods for MASC. F_1 denotes Macro-F1. MF refers to Maximization Fusion, MLR to Multi-response Linear Regression (Ting and Witten 1999), PoE to Product of Experts (Hinton 1990), MLP to Multilayer Perceptron fusion, PV to Plurality Voting, and SCDE to Sentiment Classification Decision Ensemble.

Conclusion

In this paper, we propose DEQA for MABSA and its subtasks, addressing the challenge of bridging the gap between text and visual modalities without pre-training. In DEQA, we frame MABSA as a multi-turn question-answering problem, where text, image, and description are input into separate experts in various combinations. By integrating these expert outputs within a multi-expert ensemble decision-making approach, our method generates the final predictions, achieving state-of-the-art performance. Furthermore, our framework substantially outperforms GPT-4o and other (multimodal) large language models.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62406151).

References

Bridle, J. 1989. Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters. In Touretzky, D., ed., *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann.

Chen, C.-F. R.; Fan, Q.; and Panda, R. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 357–366.

Chen, G.; Tian, Y.; and Song, Y. 2020. Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 272–279. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Egiazarian, V.; Panferov, A.; Kuznedelev, D.; Frantar, E.; Babenko, A.; and Alistarh, D. 2024. Extreme Compression of Large Language Models via Additive Quantization. arXiv:2401.06118.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Hinton, G. E. 1990. 20 - CONNECTIONIST LEARNING PROCEDURES11This chapter appeared in Volume 40 of Artificial Intelligence in 1989, reprinted with permission of North-Holland Publishing. It is a revised version of Technical Report CMU-CS-87-115, which has the same title and was prepared in June 1987 while the author was at Carnegie Mellon University. The research was supported by contract N00014-86-K-00167 from the Office of Naval Research and by grant IST-8520359 from the National Science Foundation. In Kodratoff, Y.; and Michalski, R. S., eds., *Machine Learning*, 555–610. San Francisco (CA): Morgan Kaufmann. ISBN 978-0-08-051055-2.

Hu, M.; Peng, Y.; Huang, Z.; Li, D.; and Lv, Y. 2019. Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification. arXiv:1906.03820.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991.

Ju, X.; Zhang, D.; Xiao, R.; Li, J.; Li, S.; Zhang, M.; and Zhou, G. 2021. Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4395–4405. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Khan, Z.; and Fu, Y. 2021. Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, 3034–3042. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F.; Burges, C.; Bottou, L.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Krogh, A.; and Hertz, J. 1991. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4.

Lafferty, J.; McCallum, A.; Pereira, F.; et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, 3. Williamstown, MA.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461.

Ling, Y.; Yu, J.; and Xia, R. 2022. Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2149–2159. Dublin, Ireland: Association for Computational Linguistics.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Moon, S.; Neves, L.; and Carvalho, V. 2018. Multimodal Named Entity Recognition for Short Social Media Posts. arXiv:1802.07862.

Peng, T.; Li, Z.; Wang, P.; Zhang, L.; and Zhao, H. 2024. A Novel Energy Based Model Mechanism for Multi-Modal Aspect-Based Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18869–18878.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Nakov, P.; and Zesch, T., eds., *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35. Dublin, Ireland: Association for Computational Linguistics.

Prechelt, L. 1998. *Early Stopping - But When?*, 55–69. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-49430-0.

Qi, Z.; Fang, Y.; Zhang, M.; Sun, Z.; Wu, T.; Liu, Z.; Lin, D.; Wang, J.; and Zhao, H. 2023. Gemini vs GPT-4V: A Preliminary Comparison and Combination of Vision-Language Models Through Qualitative Cases. arXiv:2312.15011.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. *nature*, 323(6088): 533–536.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Ting, K. M.; and Witten, I. H. 1999. Issues in stacked generalization. *Journal of artificial intelligence research*, 10: 271–289.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wu, H.; Cheng, S.; Wang, J.; Li, S.; and Chi, L. 2020a. Multimodal Aspect Extraction with Region-Aware Alignment Network. In Zhu, X.; Zhang, M.; Hong, Y.; and He, R., eds., *Natural Language Processing and Chinese Computing*, 145–156. Cham: Springer International Publishing. ISBN 978-3-030-60450-9.

Wu, H.; Cheng, S.; Wang, J.; Li, S.; and Chi, L. 2020b. Multimodal Aspect Extraction with Region-Aware Alignment Network. In Zhu, X.; Zhang, M.; Hong, Y.; and He, R., eds., *Natural Language Processing and Chinese Computing*, 145–156. Cham: Springer International Publishing. ISBN 978-3-030-60450-9.

Wu, Z.; Zheng, C.; Cai, Y.; Chen, J.; Leung, H.-f.; and Li, Q. 2020c. 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*, MM '20, 1038–1046. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Yan, H.; Dai, J.; ji, T.; Qiu, X.; and Zhang, Z. 2021. A Unified Generative Framework for Aspect-Based Sentiment Analysis. arXiv:2106.04300.

Yang, L.; Na, J.-C.; and Yu, J. 2022. Cross-Modal Multitask Transformer for End-to-End Multimodal Aspect-Based Sentiment Analysis. *Information Processing & Management*, 59(5): 103038.

Yang, X.; Wu, W.; Feng, S.; Wang, M.; Wang, D.; Li, Y.; Sun, Q.; Zhang, Y.; Fu, X.; and Poria, S. 2024. MM-InstructEval: Zero-Shot Evaluation of (Multimodal) Large Language Models on Multimodal Reasoning Tasks. arXiv:2405.07229.

Yu, J.; and Jiang, J. 2019. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5408–5414. International Joint Conferences on Artificial Intelligence Organization.

Yu, J.; Jiang, J.; and Xia, R. 2020. Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 429–439.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3342–3352. Online: Association for Computational Linguistics.

Yu, Z.; Yu, J.; Fan, J.; and Tao, D. 2017. Multi-Modal Factorized Bilinear Pooling With Co-Attention Learning for Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhang, L.; Wang, S.; and Liu, B. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4): e1253.

Zhou, R.; Guo, W.; Liu, X.; Yu, S.; Zhang, Y.; and Yuan, X. 2023. AoM: Detecting Aspect-oriented Information for Multimodal Aspect-Based Sentiment Analysis. arXiv:2306.01004.