TMFN: A Target-oriented Multi-grained Fusion Network for End-to-end Aspect-based Multimodal Sentiment Analysis

Di Wang[1,2*], **Yuzheng He**[1], **Xiao Liang**[1], **Yumin Tian**[1], **Shaofeng Li**[1], **Lin Zhao**[3]
[1]School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[2]Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China
[3]School of Computer Science and Engineering, Nanjing University
of Science and Technology, Nanjing 210094, China
wangdi@xidian.edu.cn, heyuzheng@stu.xidian.edu.cn, ecoxial2012@outlook.com
ymtian@mail.xidian.edu.cn, lishaofeng@xidian.edu.cn, linzhao@njust.edu.cn

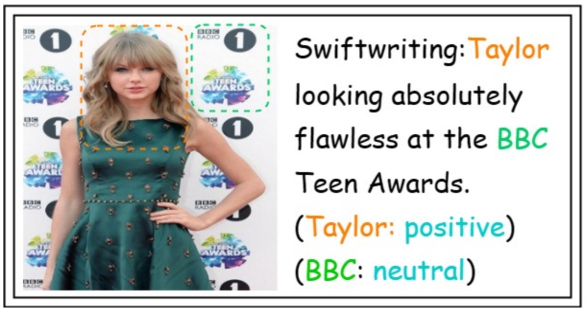
Abstract

End-to-end multimodal aspect-based sentiment analysis (MABSA) combines multimodal aspect terms extraction (MATE) with multimodal aspect sentiment classification (MASC), aiming to simultaneously extract aspect words and classify the sentiment polarity of each aspect. However, existing MABSA methods have overlooked two issues: (i) They only focus on fusing image regional information and textual words for two subtasks of MABSA. Whereas, MATE subtask relies more on global image information to assist in obtaining the quantity and attributes of aspects. Ignoring the integration with global information may affect the performance of MABSA methods. (ii) They fail to take advantage of target information. Nevertheless, the fine-grained details of targets are important for classifying sentiments of aspects. To solve these problems, we propose a Target-oriented Multi-grained Fusion Network (TMFN). It fuses text information with global coarse-grained image information for MATE subtask and with fine-grained image information for MASC subtask. In addition, a target-oriented feature alignment (TOFA) module is designed to enhance target-related information in image features with target details. In such a way, image features will contain more target emotional-related information which is beneficial to sentiment classification. Extensive experiments show that our method outperforms state-of-the-art methods on two benchmark datasets.

Keywords: Multimodal Sentiment Analysis, Aspect Terms Extraction, Aspect Sentiment Classification

1. Introduction

In recent years, with the rapid development of the Mobile Internet, people often post comments (often consisting of images and texts) online to express their opinions. The analysis of these multimodal comments has a great practical effect, which can help the government get people's attitudes to certain events to make the right decisions and can also help manufacturers get users' opinions on the products to make improvements. Aspect-based multimodal sentiment analysis, which can obtain users' views on an event from multiple perspectives in a fine-grained manner, has attracted wide attention in recent years. Early aspect-based multimodal sentiment analysis methods usually have two independent tasks. The first task is multimodal aspect terms extraction (MATE) (Li and Lam, 2017; Wu et al., 2020), which aims to extract multiple aspect words in sentences from image-text pairs. The second task is multimodal aspect sentiment classification (MASC) (Yang et al., 2022a; Huang et al., 2022), which aims to classify the sentiment polarity of aspect words in image-text pairs.



Swiftwriting: Taylor looking absolutely flawless at the BBC Teen Awards.
(Taylor: positive)
(BBC: neutral)

Figure 1: An example of end-to-end MABSA task. The input is an image-text pair. The output is aspect words 'Taylor' and 'BBC' and their sentiments.

However, extracting aspect words and classifying their emotional polarity separately ignores correlations between the two tasks and may have error accumulation, as the correctness of the extracted aspect words determines the accuracy of sentiment analysis. Moreover, learning two individual models for MATE and MASC tasks has a large time and computing consumption. To this end, Ju et al. (2021) proposed the end-to-end multimodal aspect-based sentiment analysis (MABSA) task, which performs MATE and MASC simultaneously through a single model and has better application prospects.

Existing MABSA methods focus on extracting image local information and the alignment of local image information and words. For example,

[*] Di Wang is the corresponding author.

LREC-COLING 2024, pages 16187–16197
20-25 May, 2024.

multi-modal joint learning model (JML)(Ju et al., 2021) processes the two subtasks of MABSA in stages, and uses the pre-trained model Resnet(He et al., 2016) to extract fine-grained region features of images to guide the two subtasks. Cross-modal multitask transformer model (CMMT)(Yang et al., 2022b) predicts adjective-noun pairs (ANP) and calculates loss, thereby improving the expressive ability of fine-grained features of images. Aspect-oriented method (AoM)(Zhou et al., 2023) uses region fine-grained features and text features for semantic tree modeling. Dual-encoder transformer with cross-modal alignment model (DTCA)(Yu et al., 2022b) minimizes the earth mover distance (EMD) between the text features and the fine-grained image patch features and aligns them.

For MABSA task, it is crucial to integrate image global coarse-grained information and local fine-grained information. For the reason that MATE subtask relies more on global image information to assist in obtaining the quantity and attributes of aspects, while MASC subtask relies more on fine-grained image information for classifying the sentiment of each aspect word. As shown in Figure 1, we can get target such as 'Taylor' and 'BBC' from the global information of the image, and classify Taylor's sentiment polarity through the fine-grained information of the image. In addition, the details of the target will help us obtain more sentiment-related information to improve the judgment of the sentiment polarity of the aspect words. As shown in Figure 1, Taylor's facial expression provides emotional information about the smile, which will help us classify her sentiment polarity.

In this paper, we propose a Target-oriented Multi-grained Fusion Network (TMFN) model. The core of TMFN is to fuse textual words with multi-grained image information through two independent cross-modal fusion modules and enhance the importance of emotion-related detail information of targets through target-oriented feature alignment (TOFA) module. Specifically, we divide TMFN model into three layers, each of which gets features that focus on different granularity information of the image, and then calculate the losses to bring the prediction closer to the ground-truth answer. For two independent cross-modal fusion modules in the latter two layers, we introduce a dynamic gating mechanism to control the proportion of different granularity information in the image and redesign the residual structure to enhance the text information. For the TOFA module, we calculate the similarity between patches and the candidate targets, then re-weight the candidate targets, get the target enhancement information related to each patch, and then make residual connection with each patch, thus enhancing the perception of target details.

Our contribution can be summarized as follows:

- A novel MABSA method named TMFN is proposed. It utilizes different granularity of image information for MATE and MASC subtasks. By this means, the quantity and attributes of extracted aspect words will be more accurate.
- A target-oriented feature alignment module is designed to enhance the emotional-related information in image features and consequently improve the classification accuracy of sentiment polarity.
- The proposed TMFN model outperforms baselines on two benchmark datasets Twitter2015 and Twitter2017. Specifically, F1, accuracy and recall rates increased by 0.43%, 0.53% and 0.12% on the twitter2015 dataset, and 0.54%, 1.06% and 0.03% on the Twitter2017 dataset.

2. Related Work

2.1. Aspect-based Sentiment Analysis

Aspect-based sentiment analysis of textual data has been extensively studied. This task focuses on perceiving contextual semantic information related to aspect words. Early methods proposed by Ali et al. (2016); Liu et al. (2018); Ma et al. (2018); Fan et al. (2018) are mostly based on SVM, RNN, LSTM and other traditional machine learning algorithms. Later, due to the excellent performance of attention mechanism and large-scale pre-trained language models in various text-related tasks, many recent aspect-based sentiment analysis methods have focused on utilizing them to accurately identify emotional information(Hoang et al., 2019; Li et al., 2019; Zhao and Yu, 2021). Such as Li et al. (2019) use Bert as pre-train model and explore the effects of fine-tuning with self-attention mechanism. Zhao and Yu (2021) adopt external sentiment knowledge base to enhance Bert's domain knowledge, thereby improving sentiment analysis ability.

2.2. Multimodal Aspect Sentiment Classification

With the continuous development of social media in the direction of multimodality, people find that image visual information can be used as an important supplementary information to text, so sentiment analysis keeps developing towards multimodality (Yu et al., 2022a; Zhu et al., 2022).

For MASC tasks, the existing methods focus on the acquisition of high quality fine-grained image information and image-text alignment. Image-target matching network (ITM)(Yu et al., 2022a) filters out coarse-grained information that is not relevant to targets by calculating the correlation between text and the candidate regions of the image. By setting the KL divergence function, the image target is aligned with the fine-grained aspect of the given aspect word. Face sensitive image-to-emotional-text translation method (FITE)(Yang et al., 2022a) extracts facial pictures from multimodal data and send them into the face classifier to obtain the description of the facial information, and use the facial description with the highest similarity to the given aspect words as the textually supplementary information to make the final judgment. In addition to that, sequential cross-modal semantic graph model (SeqCSG)(Huang et al., 2022) uses images to generate caption and semantic graph triples as supplementary text information. These MASC methods are unable to target the MABSA task's needs to pay attention to both image global and local fine-grained information.

2.3. End-to-End Multimodal Aspect-Based Sentiment Analysis

As a new research direction, end-to-end multimodal aspect-based sentiment analysis has more extensive practical significance and has received more and more attention. Compared with other multi-modal sentiment analysis tasks, this task is more difficult and integrated.

In the past two years, many excellent methods of end-to-end multimodal analysis have been pouring out. Multi-modal joint learning model (JML)(Ju et al., 2021) proposes a two-stage method to judge the position of aspect words and the emotional polarity of aspect words in two steps, and designs a correlation detection method of pictures and texts to screen picture information. Cross-modal multitask transformer model (CMMT)(Yang et al., 2022b) extracts image features, uses image features to predict noun adjective pairs to improve image features' quality, and introduces cross-modal transformer for image text features fusion. In addition, dual-encoder transformer with cross-modal alignment model (DTCA)(Yu et al., 2022b) aligns image patch features and text features by minimizing the earth mover distance between them, so as to better integrate the two kinds of features. Currently, the latest aspect-oriented method (AOM)(Zhou et al., 2023) in this field is innovative in that it introduces a graph neural network to model the correlation information between pictures and texts, and thus detects the location of aspect words and the sentiment polarity of each aspect word by means of triplet generated by the generating model.

However, as shown in Figure 5, the above MABSA methods don't notice the following two issues: MABSA task consists of two subtasks that need to pay attention to image different granularity information to help text information extract multiple aspects and classify their sentiment polarities. In addition, it also need to align image local features and targets to enhance the perception of the targets' emotional-related details and further improve the ability to classify the sentiment polarity. So we propose an end-to-end multimodal aspect-based sentiment analysis method named Target-oriented Multi-grained Fusion Network (TMFN) which has modular fusion design and a target-oriented feature alignment (TOFA) block to solve the above two problems.

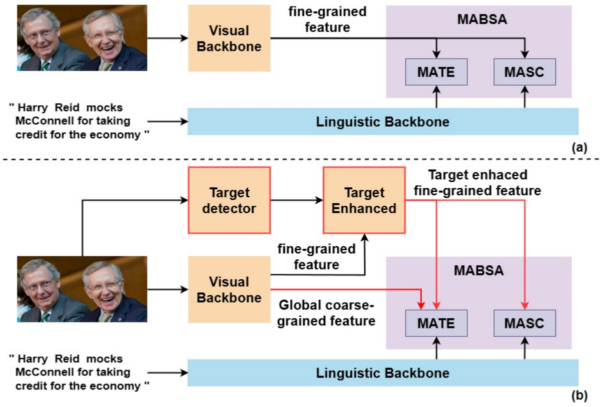

Visual Backbone
fine-grained feature
MABSA
MATE
MASC
" Harry Reid mocks McConnell for taking credit for the economy "
Linguistic Backbone
(a)
Target detector
Target Enhanced
Target enhaced fine-grained feature
fine-grained feature
Visual Backbone
Global coarse-grained feature
MABSA
MATE
MASC
" Harry Reid mocks McConnell for taking credit for the economy "
Linguistic Backbone
(b)

Figure 2: Fig (a) and (b) are flowcharts of existing models and the proposed model, respectively.

3. The Proposed TMFN Model

3.1. Overview

Task Definition. This task inputs image V and text sequence $S = \{s_1, s_2, ..., s_n\}$, where n is the number of words in the sentence, and outputs a sequence $M = \{ts_1, te_1, s_1, ..., ts_i, te_i, s_i, ..., ts_k, te_k, s_k\}$ containing all aspects of the text and their corresponding sentiment polarities, where ts_i, te_i, s_i indicates the beginning index, end index and the sentiment polarity of the i_{th} aspect word, k is the number of aspect words.

Prediction Define. In our model, We use five BIO labels {B-POS, B-NEU, B-NEG, I, O} prediction on each word in the sentence to get the output sequence M, where 'B' represents the beginning of the aspect word, 'I' indicates the end of the aspect word, and 'POS', 'NEU', and 'NEG' represent the sentiment polarity of one aspect word.

Model Preview. The proposed model architecture diagram is shown in Figure 2 , which

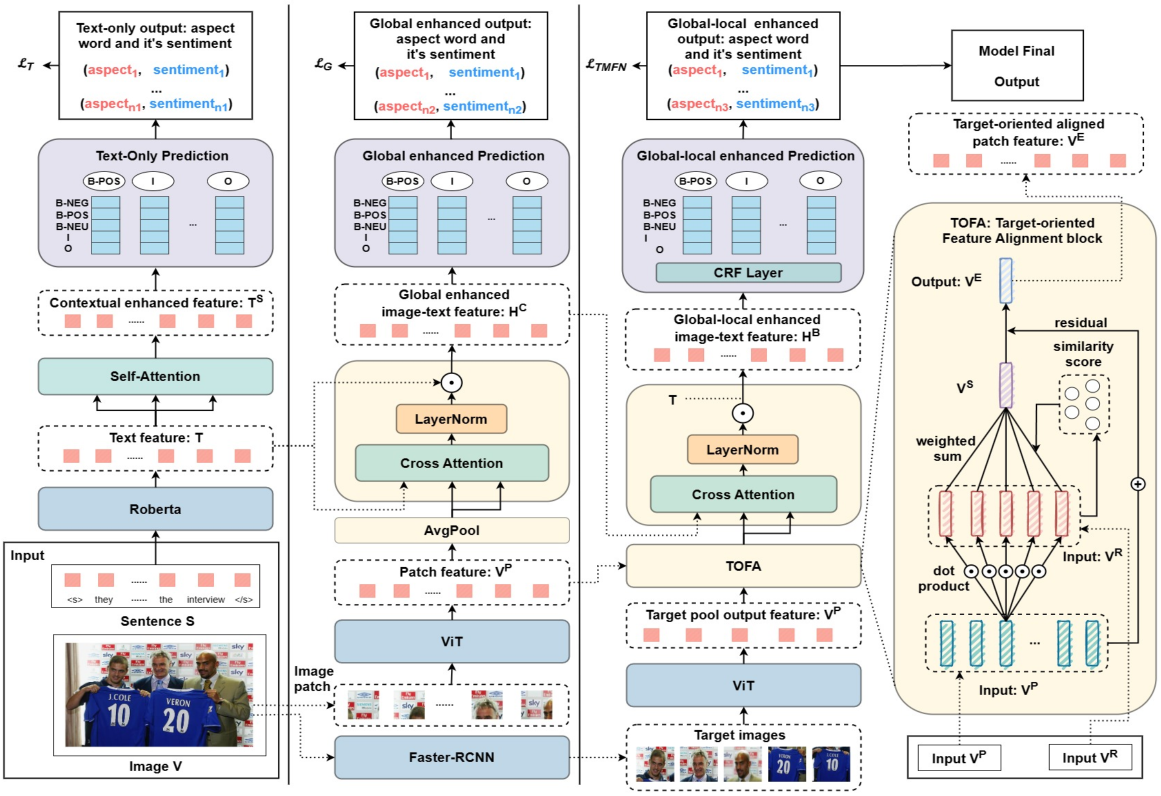


Figure 3: The overall structure diagram of the proposed TMFN model.

is divided into three layers from left to right, namely text-only layer, global enhanced layer and global-local enhanced layer. We innovatively fused text information with image different grained information by cross-modal interaction modules in the second and third layers to improve the performance of the two MABSA task's subtasks respectively. Then the TOFA module is designed to make the image information pay more attention to the emotional-related information in the target details, and further improve the performance of MASC task. In subsequent chapters, we will describe the proposed model in detail.

Feature Extractor. We adopted Roberta(Liu et al., 2019) as our text encoder. Specifically, we insert '<s>', '</s>' two special marks at the beginning and end of each input sentence S as distinctions, and then feed it into Roberta to obtain the n words embedded representation sequence $T = \{t_1, t_2, ..., t_n\}$ of text sequence S. For images, we choose ViT(Dosovitskiy et al., 2020) as our pre-train model for the reason that ViT uses patch blocks and we can obtain global coarse-grained information and preliminary regional fine-grained information more easily.

3.2. Text-only Layer

In this layer, we use text-only feature to make predictions. Specifically, we take the text feature T obtained by the textual encoder, feed it into an additional Multi-head Self-attention (MHSA)(Vaswani et al., 2017) layer to enhance the contextual awareness of feature T, and obtain the output feature $T^S = \{t^{s_1}, t^{s_2}, ..., t^{s_n}\}$, which is used for the text-only BIO labels prediction. The specific formulas are as follows:

$$T^S = MHSA(T, T, T) \tag{1}$$

$$t^{p_i} = softmax({W_t}^{\top} t^{s_i} + b_t) \tag{2}$$

$$\mathcal{L}_T = -\frac{1}{M \times N} \sum_{j=1}^{M} \sum_{i=1}^{N} CLoss(t^{p_i}, k_i) \tag{3}$$

where T^P is the prediction probability distribution of BIO labels, $W_t \in \mathbb{R}^{d \times 5}$, $\mathcal{L}_T$ is the text-only loss, M is the number of samples, N is the number of tokens in the j_{th} sample, k_i denotes the ground-truth label for the i_{th} token, $CLoss$ denotes the cross entropy loss.

The specific calculation formulas of MHSA are

Dataset	Twitter 2015					Twitter 2017				
	Pos	Neu	Neg	MA	MS	Pos	Neu	Neg	MA	MS
Train	928	1883	368	800	278	1508	1638	416	1159	733
Vaild	303	670	149	286	119	515	517	144	375	242
Test	317	607	113	258	104	493	573	168	399	263
Total	1548	3160	630	1344	501	2516	2728	728	1933	1238

Table 1: Statics of Twitter2015 and Twitter 2017 datasets. Pos: Positive, Neu: Neutral, Neg: Negative, MA: Multi aspects, MS: Multi sentiments.

as follows:

$$CATT^i(T,T,T) = softmax \\ (\frac{[W_Q^i T]^\top [W_K^i T]}{\sqrt{d/m}})[W_V^i T]^\top \tag{4}$$

$$T^S = W_c[CATT^1(T,T,T), \\ ..., CATT^m(T,T,T)] \tag{5}$$

where m is the number of MHSA heads, $\{W_Q^i, W_K^i, W_V^i\} \in \mathbb{R}^{d/m \times d}$ are query, key, value weight matrices, $W_c \in \mathbb{R}^{d \times d}$ is the weight matrix for MHSA, [,...,] means feature concatenation, and dimension d is 768.

3.3. Global Enhanced Layer

Since the MABSA task requires information of different granularity, image global information is needed to enhance the judgment of the attributes and quantity of aspect words. Therefore, in this layer, based on dynamic gating and cross-modal interaction mechanism, we use the image global information as auxiliary information to enhance text information. Specifically, we enter the whole image into ViT, and get the patch embedding feature $V^P = \{v_1^p ... v_k^p\}$ then use average pool of V^P as the global feature V^G for the whole picture, and do calculations for the global and each local feature. The specific formula is as follows:

$$V^G = avgpool(\{v_1^p ... v_k^p\}).repeat(k) \tag{6}$$

To use image global feature enhance text feature, we remove the residual connection in the Multi-head Cross-attention (MHCA) model(Vaswani et al., 2017), and then redesign a dynamic residual structure outside the module. Specifically, we use the dot product of the text-only prediction probability distribution T^P as the dynamic gate value(Yang et al., 2022b) to control the input ratio of text and image global information. We fed text feature T which is the output of Roberta and image's global pooling feature V^G into above re-designed structure, and obtain the global enhanced feature H^C. Then we use this feature to make global BIO labels prediction and get global feature loss $\mathcal{L}_G$. The specific formulas are as follows:

$$H^M = MHCA(T, V^G, V^G) \tag{7}$$

$$g_i = t^{p_i \top} t^{p_i} \tag{8}$$

$$h^{c_i} = g_i t_i + (1 - g_i) h^{m_i} \tag{9}$$

$$H^L = softmax(W_h^\top H^C + b_h) \tag{10}$$

$$\mathcal{L}_G = -\frac{1}{M \times N} \sum_{j=1}^{M} \sum_{i=1}^{N} CLoss(h^{l_i}, k_i) \tag{11}$$

where $W_h \in \mathbb{R}^{d \times 5}$, g_i is the i_{th} token's gate value, MHSA and MHCA have the same formula, just different inputs.

3.4. Global-local Enhanced Layer

Since the features of the patch blocks are not aligned with image targets and no attention is paid to the details of targets, we innovatively design a target-oriented feature alignment block (TOFA) based on the image candidate targets to solve the above problems. Firstly, we use Faster-RCNN(Girshick, 2015) as the backbone to extract image candidate targets, and set certain conditions to screen the obtained candidate targets. Specifically, we choose five candidate targets and select targets with a length or width greater than 224 pixels to avoid noise interference effects caused by too many or too small targets. The resulting candidate targets $V_1, V_1, ..., V_5$ are then fed into ViT (shared parameters with the previous chapter) to get the pooled output as the representation of each whole target, $V^R = \{v^{r_1}, v^{r_2}, ..., v^{r_5}\}$. The formula is as follows:

$$v^{r_i} = avgpool(ViT(V_i)) \tag{12}$$

Then, we take the patch feature v^{p_i} obtained in section 3.2 and the pooled features of candidate targets v^{r_i} to calculate the similarity, get the relation scores of each patch block feature for the candidate target features, and weighted sum candidate targets features by using these relation scores to obtain each patch feature's patch-target

	Methods	Twitter 2015			Twitter 2017		
		F1	P	R	F1	P	R
Text Only	SPAN(Ju et al., 2021)	53.8	53.7	53.9	60.6	59.6	61.7
	D-GCN(Ju et al., 2021)	59.4	58.3	58.8	64.1	64.1	64.2
	Roberta(Yang et al., 2022b)	63.5	61.8	65.3	66.2	65.5	66.9
	BART(Yan et al., 2021)	63.9	62.9	65.0	65.4	65.2	65.6
Multimodal	UMT+TomBERT(Ju et al., 2021)	59.8	58.4	61.3	62.4	62.3	62.4
	UMT-collapse(Yang et al., 2022b)	61.6	60.4	61.6	60.8	60.0	61.7
	UMT-Robert(Yang et al., 2022b)	63.9	61.6	66.4	66.7	65.3	68.2
	CapTrRoberta(Yang et al., 2022b)	63.2	60.6	66.1	67.3	67.1	67.4
	JML(Ju et al., 2021)	64.1	65.0	63.2	66.0	66.5	65.5
	VLP-MABSA(Ling et al., 2022)	66.6	65.1	68.3	68.0	66.9	69.2
	CMMT(Yang et al., 2022b)	66.5	64.6	68.7	68.5	67.6	69.4
	DTCA(Yu et al., 2022b)	68.4	67.3	**69.5**	**70.4**	**69.6**	**71.2**
	AoM(Zhou et al., 2023)	**68.6**	**67.9**	69.3	69.7	68.4	71.0
Our	TMFN	**69.03**	**68.43**	**69.62**	**70.95**	**70.66**	**71.23**

Table 2: Results of different methods on Twitter 2015 and Twitter 2017 for MABSA task. Bold and underlined numbers indicate the best results and bold numbers indicate the second-best results.

related feature v^{s_i}, and finally make a residual connection with each patch feature to obtain the target-oriented aligned feature $V^E = \{v^{e_1}...v^{e_k}\}$, where $k = 196$ and hidden dimension is d. The specific formulas are as follows:

$$V^S = V^P {V^R}^\top V^R \tag{13}$$

$$V^E = V^P + V^S \tag{14}$$

To utilize aligned local fine-grained information for enhancing the classification ability of the sentiment polarity of aspect words, we firstly feed global enhanced feature H^C and target-oriented aligned feature V^E into MHCA, then use dynamic gate and made a residual connection with text feature T, finally we get multimodal global-local enhanced feature H^B. The specific formulas are as follows:

$$H^O = MHCA(H^C, V^E, V^E) \tag{15}$$

$$h^{g_i} = (1 - g_i)h^{o_i} \tag{16}$$

$$H^B = H^G + T \tag{17}$$

where H^O is the output of the second MHCA layer, H^G is the gated feature of H^O and H^B is the final output feature of global-local enhanced layer.

We feed multimodal global-local enhanced feature H^B into MLP and input a standard CRF layer to predict the five BIO labels classification corresponding to each word, and get the final output sequence M. The specific formulas are as follows:

$$P(y) = \frac{exp(score(H^B, y))}{\sum_{y' \in Y_H} exp(score(H^B, y'))} \tag{18}$$

$$score(H^B, y) = \sum_{k=0}^{n} M_{y_k, y_{k+1}} + \sum_{k=0}^{n} w^{y_k} \cdot h^{b_k} \tag{19}$$

where $M_{i,j}$ is the transition matrix and represents the transition score from label i to label j, $w^{y_k} \in \mathbb{R}^{2d}$ is the weight vector for h^{b_k}. We minimize the negative log-probability of ground-truth labels as our loss function.

$$\begin{aligned} \mathcal{L}_{GARA} = -\frac{1}{M}\sum_{j=1}^{M}(score(h^{b_j}, y_j) \\ -log \sum_{y'_j \in Y_H^j} exp(score(h^{b_j}, y'_j))) \end{aligned} \tag{20}$$

where M is the number of samples, y_j denotes the jth example's ground-truth label and Y_H^j represents all possible label sequences for input tokens.

3.5. Loss Function.

We use hyper-parameters α, β to control the text-only loss $\mathcal{L}_T$ and global feature loss $\mathcal{L}_G$ and use $\mathcal{L}_{GARA}$ as our main loss to get our final loss $\mathcal{L}$. The specific formula is as follows:

$$\mathcal{L} = \mathcal{L}_{GARA} + \alpha\mathcal{L}_T + \beta\mathcal{L}_G \tag{21}$$

3.6. Experimental settings

Datasets: We use two benchmark datasets of MABSA task, Twitter2015 and Twitter2017, to evaluate the effects of TMFN model proposed in this paper. These two datasets are originally provided by (Zhang et al., 2018) for Multimodal Named Entity Recognition and labeled with sentiment polarity by (Lu et al., 2018) so that they can be used for MABSA task. The statistical results of these two datasets are shown in Table 1.

Expermental details: Our experiments are based on Roberta and ViT pre-trained models. For

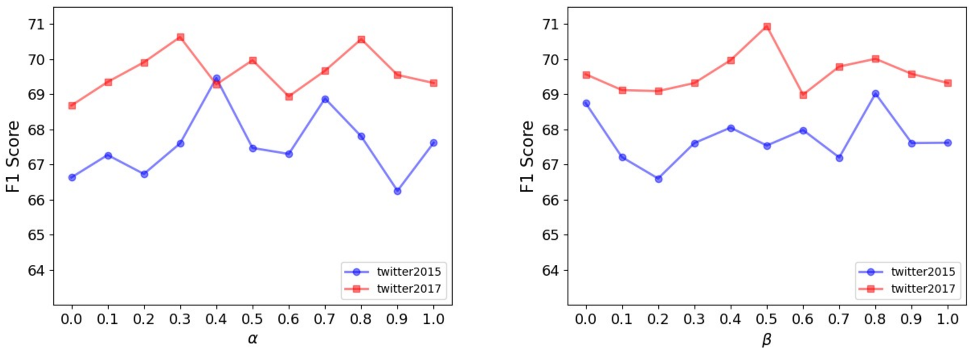


Figure 4: Hyper-parameters sensitivity experiments on Twitter2015 and Twitter2017. Set β = 1 when fine-tune α and set α = 1 when fine-tune β.

Roberta model, we use $roberta-base$, where the hidden size is 768 and the maximum sentence length is set to 60. For the vit model, we use $vit-base-patch16-224-in21k$, and the output feature dimension is consistent with the text feature. The task was implemented using PyTorch, training 25 epochs on GTX3090 with batch size 4, using the AdamW(Loshchilov and Hutter, 2017) optimizer with a learning rate of 2e-5 and warmup decay of 0.1 to update all trainable parameters. The Multi-head Self-attention and Cross-attention modules have an attention head count of 8.

Evaluation indicators: We use three metrics to evaluate the performance of TMFN model in this paper, namely Micro-F1 (F1), Precision (P) and Recall (R). Only when aspect word and it's sentiment polarity matched both can be considered correct.

For a comprehensive comparison, we selected both text-only and multimodal baseline models, as follows:

Text-only approaches for ABSA: 1) SPAN(Hu et al., 2019) first extract targets and then classify them by using Span-based Scheme. **2) D-GCN**(Chen et al., 2020a) build dependence tree among words and improved GCN so that it can adapt the category relationships of specific images in an adaptive way. **3) Roberta**(Liu et al., 2019) is based on Bert and improved the performance. **4) BART**(Yan et al., 2021) adopts encoder decoder architecture, which is a generation model based on transformer.

Multimodal approaches for MABSA: 1) UMT+TomBERT(Yang et al., 2022b) is pipeline approach which combine UMT(Yu et al., 2020) and TomBERT(Yu and Jiang, 2019a). **2) UMT-collapse**(Yu et al., 2020) which is designed for MATE task and then adapted to MABSA task uses Cross-Modal Transformer layers to fuse text and picture information. **3) UMT-Robert**(Yang et al., 2022b) replaces text encoder with Roberta from UMT-collapse. **4) JML**(Ju et al., 2021) first proposes End-to-End MABSA task and uses two step to extract aspect words and classify their corresponding sentiment polarity. **5) CapTrRoberta**(Khan and Fu, 2021) uses images to generate auxiliary sentence as supplement of original text. **6) VLP-MABSA**(Ling et al., 2022) is a pre-train model for MABSA, using BART as base model. **7) CMMT**(Yang et al., 2022b) predicts ANP pairs to improve the expression ability of image features and narrow the distance with text features. **8) DTCA**(Yu et al., 2022b) aligns image feature and text feature by minimize the Wasserstein distance. **9) AoM**(Zhou et al., 2023) models the dependencies between image and text and fuses them with GCN.

4. Experiment

4.1. Baselines

4.2. Experimental Results

In this section, we will compare the baseline models and fully demonstrate the excellent results achieved by the proposed TMFN. The results of different models are shown in Table 2.

Compared with the above baseline models, our model achieves SOTA on the three evaluation metrics(F1, P, R) of the two datasets. Among them, in the metric of P, compared with the previous DTCA[1] model with the best comprehensive performance, our model has respectively increased by 1.13% and 1.06%, and compared with the newly proposed AoM model, it has respectively increased by 0.53% and 2.26%. This fully demonstrates the effectiveness of our designed method for focusing on image local information and enhancing regional features by aligning with targets. In addition, the metric of F1 in the two datasets is 0.63%,

[1] This model also uses ViT as image encoder.

Images		
Text	Embattled Metro Councilman Dan Johnson to debate challenger John Witt , an independent .	Harry Reid mocks McConnell for tacking credict for the economy
JML	(Metro, Neu) (Dan Johnson, Neu) ✗	(Harry Reid, Pos) (McConnell, Neu) ✓
CMMT	(Metro, Neu) (Dan Johnson, Neu) (John Witt, Neu) ✓	(Harry Reid, Neu) (McConnell, Neu) ✗
TMFN	(Metro, Neu) (Dan Johnson, Neu) (John Witt, Neu) ✓	(Harry Reid, Pos) (McConnell, Neu) ✓

Figure 5: Case Analysis. JML and CMMT are comparison models. TMFN is the proposed model. Words in red font are aspect words and in blue font are sentiment words.

0.55% and 0.43%, 1.25% higher than that of DTCA and AoM respectively. We analyze the reason and speculate that because the above models may ignore to pay attention to global information and local information at the same time, which may lead to misjudgment of the number and position of aspect words. Thus, the comprehensive evaluation metric F1 is significantly affected. This also shows the effectiveness of the proposed TMFN model in integrating multi-grained information.

4.3. Analysis

Ablation Study: To further demonstrate the effects of the TMFN models we proposed, ablation experiments are designed as follows: **1) w/o TOFA** means we remove the TOFA block. **2) w/o G-enhanced** means we remove the global enhanced layer and don't use image global information to enhance text feature. We also remove the layer loss $\mathcal{L}_G$. **3) w/o L-enhanced** means we remove the Global-local enhanced layer and also remove this layer loss $\mathcal{L}_{TMFN}$. **4) w/o G-L-enhanced** means we remove the above two layers and only use text feature for prediction. **5) w/o Gate control** means we remove the dynamic gate control mechanism in TMFN model.

The results are shown in Table 3. From the form, we can see that after removing each block, each evaluation matric has decreased, which indicates the effectiveness of our proposed model. On Twitter2015, the removal of the gating mechanism has the greatest impact on the overall performance of the model, with the F1 value decreasing by 1.76%. This shows that the non-correlation between picture and text may be more obvious on Twitter2015 dataset, and more attention needs to be paid to noise control. On Twitter2017, w/o G-enhanced has the biggest effect, with F1 dropping by 2.1%, suggesting that images may provide more global information to help judge the content and quantity of aspect words on the Twitter2017 dataset. In addition, the result of w/o G-enhanced or w/o L-enhanced is mostly lower than that of w/o G-L-enhanced, which just shows that when processing MABSA tasks, we should not pay too much attention to a certain granularity of information, otherwise image noise may be introduced, and then affect the quality of text features. We should consider both the image global and local comprehensive.

Methods	Twitter 2015			Twitter 2017		
	F1	P	R	F1	P	R
TMFN	69.03	68.43	69.62	70.95	70.66	71.23
w/o TOFA	67.47	65.20	69.91	69.89	69.61	70.17
w/o G-enhanced	68.12	67.04	69.24	68.85	68.57	69.12
w/o L-enhanced	67.92	66.30	69.62	69.22	68.21	70.26
w/o G-L-enhanced	68.57	67.93	69.24	68.97	68.34	69.61
w/o Gate Control	67.27	66.47	68.08	69.15	69.01	69.29

Table 3: Results of ablation experiments.

Hyper-parameters Fine Tuning: We set up related experiments with hyper-parameters α and β to explore their influence on the experimental results. Specifically, on the two datasets, we conducted relevant experiments by fixing one of the two parameters to 1 respectively, and then increase the other parameter from 0 to 1 by 0.1, and the results are shown in Figure 3. Finally, we get result by using $(1, 0.8)$ hyper-parameter pairs in Twitter 2015 dataset and $(1, 0.5)$ in Twitter 2017 dataset as the final result of our proposed DTCA model. In addition, subsequent ablation experiments also fixed the hyper-parameter to this optimal value.

Case Study: To better demonstrate the effect of TMFN model, we analyzed two cases as shown in Figure 4. We choose CMMT and JML model to compare. In the first example, JML lacks focus on the global information and therefore fails to identify the correct number of aspect terms. In the second example, we designed the TOFA module to align image local feature with target and enhance the perception of target emotional-related details, which has certain advantages, so that we can better grasp the sentiment polarity of Harry Reid.

5. Conclusion

In this paper, we propose a TMFN method for MABSA. It fuses global image information and local fine-grained image information with text to improve the performance of MATE and MACS subtasks in MABSA. In addition, since the details of targets in images contain a lot of emotional-related information, we designed a TOFA module to align and enhance image regional features with sentiment information of targets, so as to improve the accuracy of the MASC subtask. Experiments on two benchmark datasets show that our method outperforms state-of-the-art results.

In future work, we will try to strengthen fine-grained text information for MABSA and apply our model to other multimodal tasks that require attention to different granularity of information.

6. Acknowledgements

This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0117103, in part by the National Natural Science Foundation of China under Grants 62072354, 62172222 and 62072355, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23084, in part by the Key Research and Development Program of Shannxi Province under Grant 2024GX-YBXM-039, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2024JC-YBQN-0340, in part by the Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, and in part by a grant from the Youth Innovation Team of Shaanxi Universities under Grant 2023-CX-TD-08.

7. Bibliographical References

Farman Ali, Kyung-Sup Kwak, and Yong-Gi Kim. 2016. Opinion mining based on fuzzy domain ontology and support vector machine: A proposal to automate online review classification. *Applied Soft Computing*, 47:235–250.

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

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

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

Chuang Fan, Qinghong Gao, Jiachen Du, Lin Gui, Ruifeng Xu, and Kam-Fai Wong. 2018. Convolution-based memory network for aspect-based sentiment analysis. In *The 41st International ACM SIGIR conference on research & development in information retrieval*, pages 1161–1164.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

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.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics*, pages 187–196.

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

Yufeng Huang, Zhuo Chen, Wen Zhang, Jiaoyan Chen, Jeff Z Pan, Zhen Yao, Yujie Xie, and Huajun Chen. 2022. Aspect-based sentiment classification with sequential cross-modal semantic graph. *arXiv preprint arXiv:2208.09417*.

Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. 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.

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.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting bert for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2886–2892.

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

Fei Liu, Trevor Cohn, and Timothy Baldwin. 2018. Recurrent entity networks with delayed memory update for targeted aspect-based sentiment analysis. *arXiv preprint arXiv:1804.11019*.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Yukun Ma, Haiyun Peng, Tahir Khan, Erik Cambria, and Amir Hussain. 2018. Sentic lstm: a hybrid network for targeted aspect-based sentiment analysis. *Cognitive Computation*, 10:639–650.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat*, 1050(20):10–48550.

Bing Wang, Liang Ding, Qihuang Zhong, Ximing Li, and Dacheng Tao. 2022. A contrastive cross-channel data augmentation framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2204.07832*.

Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020. 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.

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

Hao Yang, Yanyan Zhao, and Bing Qin. 2022a. 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.

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

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

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

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics.

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.

Zhewen Yu, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2022b. Dual-encoder transformers with cross-modal alignment for multimodal aspect-based sentiment analysis. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 414–423.

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.

Anping Zhao and Yu Yu. 2021. Knowledge-enabled bert for aspect-based sentiment analysis. *Knowledge-Based Systems*, 227:107220.

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

Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. 2022. Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia*.