Abstract

Multi-modal aspect-based sentiment classification (MABSC) is an emerging classification task that aims to classify the sentiment of a given target such as a mentioned entity in data with different modalities. In typical multi-modal data with text and image, previous approaches do not make full use of the fine-grained semantics of the image, especially in conjunction with the semantics of the text and do not fully consider modeling the relationship between fine-grained image information and target, which leads to insufficient use of image and inadequate to identify fine-grained aspects and opinions. To tackle these limitations, we propose a new framework SeqCSG including a method to construct sequential cross-modal semantic graphs and an encoder-decoder model. Specifically, we extract fine-grained information from the original image, image caption, and scene graph, and regard them as elements of the cross-modal semantic graph as well as tokens from texts. The cross-modal semantic graph is represented as a sequence with a multi-modal visible matrix indicating relationships between elements. In order to effectively utilize the cross-modal semantic graph, we propose an encoder-decoder method with a target prompt template. Experimental results show that our approach outperforms existing methods and achieves the state-of-the-art on two standard datasets MABSC. Further analysis demonstrates the effectiveness of each component and our model can implicitly learn the correlation between the target and fine-grained information of the image.

1 Introduction

Multi-modal aspect-based sentiment classification (MABSC) is an emerging task aiming to classify the sentiment of a given target such as a mentioned entity in data with different modalities. Specifically, given the text-image pair and the target, the goal of MABSC is to identify the target’s sentiment polarities, as shown in Figure 1. MABSC based on tweets is more challenging than text-based aspect sentiment classification and multi-modal sentiment classification for the following reasons: tweets are shorter and carry less information; tweet-related visual modal information may be irrelevant or noisy; tweet-related visual scenes are more complex than ordinary images.

Figure 1: Examples of MABSC. Given an image, a tweet and some targets, our goal aims to predict the sentiment polarity of each target.

Recent years have witnessed increasing attentions on MABSC task and many methods are proposed for this challenging task. Some studies use image caption and fuse caption and tweet to achieve model alignment. Other works model aspects, opinions and their alignment through task-specific visual language pre-training (VLP-MABSA). Although these works achieved alignment between images and texts to a certain extent, they only utilize coarse-grained information from images, such as features of the entire image, which is insufficient for correctly classifying two tasks with the same image-text pair while different targets and sentiments, as shown in Figure 1. Therefore, we argue that fine-grained information from image and text considering the target in the MABSC task should be crossly utilized.

To achieve this, in this paper, we propose to construct a cross-modal semantic graph for each image-text pair of MABSC task and represent it as a sequence with a visible matrix in order to obtain a high-level structured representation of the visual context. Specifically, we extract elements of the semantic graph from text and image, including tokens from text and image caption, and image-level knowledge such as relationship between fine-grained images and objects and the relationship between objects. Then we transform all elements of the semantic graph into a sequence, and construct the structure of the semantic graph through a multi-modal visible matrix indicating the connections between different elements in the semantic graph sequence. Meanwhile, we built a manual prompt template for the target, which aims to guide the model to build the connection between the target and other information. Inspired by
the recent success of sequence-to-sequence models in VLP-MABSA (Ling, Yu, and Xia 2022), we input the sequential multi-modal semantic graph into a sequence encoder, and construct a prompt template according to the target as the input of the decoder, converting the generation problem into a classification problem, i.e. taking the outputs of specific locations from the decoder as input for final classification.

To prove the effectiveness of our approach, we experimentally evaluate the model on two benchmarks of MABSC, Twitter2015 (Zhang et al. 2018) and Twitter2017 (Lu et al. 2018). Results show that our approach achieves better results. Specifically, compared to the strongest baseline VLP-MABSA, our approach has a 0.8% improvement in accuracy and a 1.6% improvement in the f1-score on Twitter2015. On Twitter2017, the improvements are 1.1% and 1.9%, respectively, proving that sequential cross-modal semantic graph with the visible matrix is effective for MABSC. More importantly, our model can implicitly learn the correlation between the target and fine-grained information of the image.

In summary, our main contributions are as follows:

- We propose a sequential cross-modal semantic graph construction method for MABSC, which could help crossly utilize fine-grained information from image and text in one task.
- We propose an encoder-decoder method with prompt template that could effectively utilize the sequential cross-modal semantic graph considering target in MABSC.
- We perform comprehensive experiments and extensive analysis on two datasets for MABSC illustrating that that our model can effectively and robustly model the multi-modal representations of descriptive texts and images and achieves state-of-the-art results.

2 Related Work

2.1 Text-based Target-oriented Sentiment Classification

Text-based target-oriented sentiment classification aims to predict the sentiment polarities of the target, which is a mention entity in text. This task benefits from a better aspect-based text representation. Therefore, an amount of neural network-based models were proposed to deal with challenges. For example, graph-based models conduct graph convolutional operations on dependency trees to encode the semantic information, and attention-based models focus on interacting between aspect terms and context tokens.

Recently, benefiting from the rapid development of Transformers (Vaswani et al. 2017), pre-trained language models (PLMs) have emerged because they have strong language representation and understanding ability. They can get more effective information from language encoders. For example, Dai et al. (2021) leveraged RoBERTa to reconstruct dependency trees. Yan et al. (2021) proposed a unified generative framework that achieves highly competitive performance.

2.2 Multi-modal Sentiment Classification

The goal of multi-modal sentiment classification is to discover the sentiment expressed in multi-modal samples. Inspired by the framework of multi-task learning, Yu et al. (2021b) proposed the task of multi-modal joint training and learning multi-modal and unimodal representation of sentiment analysis; Yang, Xu, and Gao (2020) extended the standard pre-trained BERT model to cross-modal scenarios and proposed a multi-modal BERT for sentiment analysis; Wu et al. (2022) designed a multi-modal emotion analysis model based on multi-head attention to capturing multi-modal emotion fusion; Keswani et al. (2020) designed a method, which used BERT’s multi-modal Bitransformer (MMBT) and ResNet to model text and visual features. With the development of multi-modal PLMs, there are also some existing works that use ViLT (Kim, Son, and Kim 2021) and LXMERT (Tan and Bansal 2019) as the backbone for multi-modal sentiment analysis.

2.3 Multi-modal Aspect-based Sentiment Classification

MABSC is a new field that combines the multi-modal sentiment analysis and the target-oriented sentiment classification. The earlier work is TomBERT (Yu and Jiang 2019), which introduces two multi-modal tweet datasets with target annotations, as well as a target-oriented multi-modal BERT model. TomBERT builds on top of the baseline BERT architecture by adding target-sensitive visual attention and more self-attention layers to capture cross-modal dynamics. EF-CapTrBERT (Khan and Fu 2021) is a work that optimizes on TomBERT, but transforms image information into image caption, and then fuses the information of the two modes. Unlike TomBERT, this model does not modify the baseline BERT model, nor does it add an additional self-attention layer. The work most related to ours is VLP-MABSA (Ling, Yu, and Xia 2022), which is a task-specific vision-language pre-training framework for MABSC. It is a unified multi-modal encoder-decoder architecture for all the pre-training and downstream tasks. In contrast to VLP-MABSA, we propose a prompt guided translation of problems into classification problems under the generation-based paradigm. In addition, we propose a method of parsing sequential cross-modal semantic graph from the image driven by visible matrix, and then interacting with text information to extract more fine-grained information from the image and establish the relationship between fine-grained information and the target.

3 Method

3.1 Task Formulation

Given a set of multi-modal samples $M$, each sample $m \in M$ consists of a sentence $s$, an image $v$, an opinion target $t$, and a label $y$, that is $m = \{s, v, t, y\}$. $s = (w_1, w_2, w_3, ..., w_n)$, where $n$ is the number of words. The opinion target $t$ is a mentioned entity in $s$. And there are three kinds of lables that $y \in \{negative, neutral, positive\}$. Multi-modal aspect-based sentiment classification can be stated as follows: given $M$, the goal is to learn a target-oriented sentiment classifier so that it can correctly predict the sentiment labels for opinion targets in unseen samples. For example, given a tweet like
“On TMZ: Golden State Warriors Rage After NBA Championship, Boozing, Dancing, Screaming! ” and the relevant image, the model should be able to predict “Positive” for the target “Golden State Warriors” and “Neutral” for the target “NBA”, as shown in Figure 1.

3.2 Overview

Given a sample \( m = \{s, v, t, y\} \), for the input image \( v \in \mathbb{R}^{3 \times H \times W} \), we first apply the captioning transformer \( \text{Khan and Fu [2021]} \) to turn the image \( v \) into corresponding image caption \( C(v) \). Meanwhile, a scene graph is extracted from the image \( v \) via scene graph generation \( \text{Tang et al. [2020]} \). Two types of triples are involved in the scene graph: (1) \( \langle h, r, t \rangle \), where \( h, t \) are textual entities and \( r \) is a relation predicate; (2) \( \langle \text{img}, \text{image of}, t \rangle \), where \( \text{img} \) is a special token, which represents the corresponding subimage (i.e., part of the entire image), “image of” is a temporary relation. We serialize these triples linearly \( \text{Chen et al. [2022a,b], Yin et al. [2020]} \) to compose the triple sentences which are separated by special token \( \langle ts \rangle \). Our cross-modal semantic graph elements consist of the tweet text, the caption and the scene graph. We transform all elements of the semantic graph into a sequence. Inspired by K-BERT \( \text{Liu et al. [2020]} \), we designed a multi-modal visible matrix to construct the structure of the semantic graph, as discussed in Sec. 3.3. Strictly speaking, the encoder of our sequence-to-sequence model consists of the triple sentence, caption and tweet, and the decoder side contains the manual prompt template composed of the target. The input construction as well as the training and prediction process of the model are elaborated in Sec. 3.4.

3.3 Sequential Cross-modal Semantic Graph Construction

Semantic Graph Element Extraction. Our semantic graph elements contain three parts: scene graph, caption and tweet text. The tweet text is the original input, and the caption is obtained by the captioning transformer in order to capture the global information of the image. As for the scene graph, unlike previous works that use object representations extracted from the image as the visual knowledge source \( \text{Chen et al. [2020], Li et al. [2020], Lu et al. [2019], Su et al. [2020], Yu et al. [2021a]} \), we utilize a pre-trained scene graph generator to extract a scene graph that consists of the Recall@5 \( \text{(subject, predicate, object)} \) triplets to represent the object-level image context \( \text{Tang et al. [2020]} \), e.g., \( \langle \text{car, behind, man} \rangle \), as well as Recall@5 \( \text{(img), image of, object)} \) triplets to represent the relation between the subimage and the object. Particularly, \( \langle \text{img} \rangle \) is a special token, which represents the relevant subimage.

Element-to-Sequence Transformation. This module unifies the dynamic multi-modal semantic graph generated from an image into a textual format. Our motivation is to represent the image in a form that the transformer already understands: a sequence of continuous and interconnect embeddings. As introduced before, two types of triples are extracted from the image. Specifically, we convert each triple in KG into a serialized sentence separated by “;”, then concatenate all those serialized triples with the special token \( \langle ts \rangle \) to construct final triple sentences. For example, given the scene graph: \( \langle \text{train has seat, (person, watching, man), (img1, image of, train), (img2, image of, person), (img3, image of, man) \rangle} \) we convert the above triples into following serialized form:

\[
\langle s \rangle \text{train, has, seat \langle ts \rangle ... \langle img, image of, man/\langle s \rangle}. \quad (1)
\]

Finally, we concatenate the caption, tweet text and triple sentence with a special token \( \langle /s \rangle \) to form the encoder input of our sequence-to-sequence model.

Semantic Graph Structure Construction. We represent a set of triples \( T_{in} \) as a sequence of tokens. Although the current serialized triple sentence already contains a lot of information about the triplet, the serialization method destroys the structure of the triplet itself and the implicit information contained between entities. In particular, the risk raised with knowledge is that it can lead to a change in the meaning of
the original sentence. Inspired by K-BERT (Liu et al. 2020) which builds a relatedness matrix to indicate the relationship between relevant entity tokens in KGs, we want the model to enhance the internal connection within the triplets meanwhile extract more valid information from the same entity in the serialized triple sentence. Concretely, (1) for input triple sentence in the encoder, we require that everything inside each triplet be visible. The shared entities within various triples are visible to each other while the rest of the triple sentence is invisible. This alleviates the noise from irrelevant information and ensures that the implicit information between entities is modeled; (2) the tweet, caption and other special tokens in the encoder should be visible to each other, so that the text information can interact with the triplet information extracted by the image. To some degree, the visible matrix M contains the structural information of the triple sentence. The visible matrix M is defined as Eq. (2),

\[
M_{ij} = \begin{cases} 
1 & \text{if } w_i \in K \text{ or } w_j \in K, \\
1 & \text{if } (w_i \in S \cup C) \text{ or } (w_j \in S \cup C), \\
1 & \text{if } (w_i \in e_1) \cap (w_j \in e_2) \cap (e_1 = e_2), \\
0 & \text{otherwise}, 
\end{cases}
\]

where \(w_i, w_j\) are tokens in sentences; \(e_1, e_2\) are entities; \(K\) is special tokens; \(S\) denotes the tweet text and \(C\) denotes the image caption.

Transformer layer includes a self-attention module and a position-wise feed-forward network. Suppose the input of self-attention module is \(H = [s_1, ..., s_n] \in \mathbb{R}^{n \times d}\) with the \(i^{th}\) row as the \(d\) dimensional hidden state for the \(i^{th}\) element. The self-attention operation is

\[
Q = HW_Q, K = HW_K, V = HW_V, 
\]

\[
A = \frac{QK^T}{\sqrt{d}}, \text{Attn}(H) = \text{softmax}(A)V, 
\]

where \(W_Q \in \mathbb{R}^{d \times d_Q}, W_K \in \mathbb{R}^{d \times d_K}, W_V \in \mathbb{R}^{d \times d_V}\) is the projection matrix to generate the query, key, and value representation of \(H\); \(A\) is the matrix capturing similarity between the query and the key.

Different from the vanilla self-attention module, we make encoder transformer layer aware of the relatedness between elements defined in \(M\) in the self-attention module that

\[
A_{ij} = \frac{M_{ij} \times (h_iW_Q)(h_jW_K)^T}{\sqrt{d}} + (1 - M_{ij}) \ast \delta, 
\]

\[
\text{Attn}(h_i) = \sum_{j=1}^{n_s} \text{softmax}(A_{ij}) \times (e_jW_V), 
\]

where \(\delta\) is a large negative number to make values after the softmax function \(\text{softmax()}\) near to 0.

3.4 Model Architecture

Based on the seq2seq architecture, we employ the generative model, which consists of two components, to classify the outputs of specific locations for aspect-based sentiment classification.

Encoder. The input of encoder is composed of serialized cross-modal semantic graph, caption and tweet text. For sentence \(X\) in the encoder, we tokenize it into a sequence of tokens \(X = \{x_1, x_2, ..., x_n\}\). The encoder is to encode \(X\) into the hidden representation space as a vector \(H_{en}\),

\[
H_{en} = \text{Encoder}(X), 
\]

where \(H_{en} \in \mathbb{R}^{n \times d}\) and \(d\) is the hidden state dimension.

The function of the embedding layer is to convert the input sentence into an embedding representation that can be fed into the Transformers. Similar to BERT, the embedding representation of our model is the sum of three parts: token/image embedding, position embedding and type embedding, but differs in that the input of the model’s embedding layer contains image and text information rather than a token sequence. Therefore, how to embed our input while retaining its structural information is the key to our model.

Considering the input of multi-modal information, token/image embedding distinguishes input tokens. For text tokens, it is consistent with BERT, and the vocabulary provided by Facebook BART is adopted in this paper. Each token in the sentence tree is converted to an embedding vector of dimension \(H\) via a trainable lookup table. For image tokens, we use ResNet to encode them and transform them into an embedding vector of dimension \(H\) through a linear layer. Specifically, To avoid having an excessive number of arguments affect the performance of the especially stacked model, during training we set ResNet’s parameters to freeze, leaving only one linear layer to learn.

Following ViLT, it is easy to set the type embedding. We set the image token embedding as 1 and the text token embedding as 0. For position embedding, without position embedding, it will be equivalent to a bag-of-word model, resulting in a lack of structural information (i.e., the order of tokens). Therefore, we followed the position embedding inherent in BART model to encode the position information.

Decoder. The decoder part takes the encoder output \(H_{en}\) and previous decoder output \(y_1, y_2, ..., y_{t-1}\) as inputs and the decoder \(y_{t}, y_{t+1}\) indicates the token indexes. Existing studies (Chen et al. 2022a; Liu et al. 2021) have shown that answer engineering has a strong influence on the performance of prompt-tuning. Different from the multi-modal sentiment classification task, the classification basis of the MABSC task does not only depend on the information of text and image. Note that the sentiment orientation will vary greatly depending on the target when dealing with the interaction and fusion of text and image information. For example, given the tweet text “Sergio Ramos chosen as the best player of UCL final.” and its corresponding image, the sentiment tendency of “Sergio Ramos” is “Positive” but “UCL” is “Neutral”.

To address the above issues, we propose to convert the target information in the input into a prompt template. Specifically, we convert the target to the form of “target is \langle\text{mask}\rangle” as a simple template is used to establish the connection between the target and sentiment orientation.

Taking “Sergio Ramos chosen as the best player of UCL final.” as an example, the input content of the encoder remains the same and is composed of three parts: serialized
We transform the target “Sergio Ramos” in this sentence into the form of “Sergio Ramos is (mask)”, input it to the decoder end, and then through the 〈mask〉 token of embedding, combined with MLP for sentiment classification. The experimental results also show that this method is simple but effective for aspect-based sentiment classification tasks.

Then for each target x_{in}, let the manipulation \( X_{\text{prompt}} = T(x_{in}) \) be a masked language modeling (MLM) input which contains one 〈mask〉 token. In this way, we can treat our task as an MLM, and model the probability of predicting class \( y \in \mathcal{Y} \) as:

\[
p(y | H_{[m]}) = \text{softmax} \left( \theta_{\text{LinearDropout}} \left( H_{[m]} \right) \right),
\]

where \( H_{[m]} \) is the hidden vector of 〈mask〉 dimension. \( \theta_{\text{Linear}} \in \mathbb{R}^{3 \times 768} \), and is learned by backpropagation. We learn \( \theta_{\text{Linear}} \) by fine-tuning the BART alongside Eq.8 using the standard cross-entropy loss.

## 4 Experiment

### 4.1 Experiment Setting

**Downstream Datasets.** We adopt two benchmark datasets annotated by [Yu and Jiang (2019)](https://example.com), namely Twitter2015 and Twitter2017 for MABSA task to evaluate our model. The statistics of the two datasets are shown in Table 1. Twitter2015 and Twitter2017 consist of multi-modal tweets, where each multi-modal tweet consists of the target, an image posted alongside the tweet text, targets within the tweet text, and the sentiment of each target. Each target is given a label from the set \{negative, neutral, positive\}, and the task is a standard multi-class classification problem.

**Implement Details.** We employ BART [Lewis et al. (2020)](https://example.com), a denoising and simple encoder-decoder PLM, as our framework. Specifically, we fix all the hyper-parameters after tuning them on the development set, and the downstream tasks are fine-tuned for 30 epochs. The batch size is set to 16; the learning rate is set to 2e-5; the hidden size of our model is set to 768, which is the same as standard BART. Note that we freeze the ResNet parameters to decrease the learnable parameters hence avoid overfitting. We implement all the models with PyTorch, and run experiments on a RTX3090 GPU.

**Evaluation Metrics.** We evaluate our model on MABSA task and adopt Accuracy (Acc) and Macro-F1 score (F1) as the evaluation metrics to measure the performance.

### 4.2 Baseline

In this section, we adopt three types of baselines.

1) **Image Only.** We select Res-Target, which directly applies cross-modal attention to ResNet input features as the language features without any extra modifications.

2) **Text Only.** We select BERT model as well as BERT+BL, which is BERT with another BERT layer stacked on it, and MGAN [Fan, Feng, and Zhao (2018)](https://example.com) which uses a multi-grain attention network for aspect understanding. Moreover, as a strong baseline, BERT-Pair-QA [Sun, Huang, and Qiu (2019)](https://example.com) obtains comparable performance on SemEval-2014 Task 4. Furthermore, we take BART as a baseline, which only takes text and target as input.

3) **Text and Image.** We compare with TomBERT [Yu and Jiang (2019)](https://example.com), which employs BERT to capture cross-modal dynamics, and CapTrBERT [Khan and Fu (2021)](https://example.com), which translates the image to a caption as an auxiliary sentence for sentiment classification, and JML [Ju et al. (2021)](https://example.com), which is a multi-task learning approach proposed recently with the auxiliary cross-modal relation detection task. We also compare with VLP-MABSA [Ling, Yu, and Xia (2022)](https://example.com), which proposed a task-specific Vision Language Pre-training framework for MABSA. Furthermore, we take Multi-BART as a strong baseline, considering that our model is a sequence-to-sequence model and we combine text, caption, and target as input.

### 4.3 Main Result

Table 2 shows the results of different methods for MABSA on Twitter2015 and Twitter2017 datasets. As we can see from the table, the performance of the image-based methods and the text-based methods are lower than the combination of the two modalities. In particular, the experimental results of the image-based methods are much lower than those of the multi-modal methods, while the experimental results of the text-based methods are not different from those of the multi-modal methods. This also indicates to some extent that there is still a lot of room for improvement in the processing of visual features and the interaction between the two modalities.

From the experimental results, we can observe that Multi-BART can also achieve good performance, even better than some multi-modal methods, which proves the superiority of our basic framework. For multi-modal methods, the performance of VLP-MABSA is better than previous methods, mainly because the model is designed for pre-training tasks under specific tasks, which is very helpful for alignment and interaction between texts and images.

Among all the methods, our proposed method performs significantly better than other baseline models on all benchmarks. Specifically, it remarkably outperforms the second-best VLP-MABSA system on Twitter2015 and Twitter2017, surpassing VLP-MABSA by 1.2 and 1.4 absolute percentage points on F1-score, respectively. It is 0.7 and 0.8 ab-
Table 2: The property prediction performance of our method (SeqCSG), compared with image-only (first group), text-only (second group) and multi-modal methods (third group) baselines on Twitter2015 and Twitter2017 datasets.

| Modality     | Method                                      | Twitter2015 Acc | Twitter2015 Macro-F1 | Twitter2017 Acc | Twitter2017 Macro-F1 |
|--------------|---------------------------------------------|-----------------|----------------------|-----------------|----------------------|
| Visual       | Res-Target (He et al. 2016)                 | 59.9            | 46.5                 | 58.6            | 54.0                 |
|              | MGAN (Fan, Feng, and Zhao 2018)            | 71.2            | 64.2                 | 64.8            | 61.5                 |
|              | BERT (Devlin et al. 2019)                  | 74.3            | 70.0                 | 68.9            | 66.1                 |
|              | BERT+BL (Devlin et al. 2019)               | 74.3            | 70.0                 | 68.9            | 66.1                 |
|              | BERT-Pair-QA (Sun, Huang, and Qiu 2019)    | 74.4            | 67.7                 | 63.1            | 59.7                 |
|              | BART (Lewis et al. 2020)                   | 76.0            | 67.6                 | 69.5            | 67.0                 |
| Text         | Res-MGAN                                    | 71.7            | 63.9                 | 66.4            | 63.0                 |
|              | Res-BERT+BL                                 | 75.0            | 69.2                 | 69.2            | 66.5                 |
|              | mPBERT (CLIS) (Yu and Jiang 2019)          | 75.8            | 71.1                 | 68.8            | 67.1                 |
|              | TomBERT (Yu and Jiang 2019)                | 77.2            | 71.8                 | 70.5            | 68.0                 |
|              | CapTrBERT (Khan and Fu 2021)               | 78.0            | 73.2                 | 72.3            | 70.2                 |
|              | JML-MASC (Yu et al. 2021)                  | 78.7            | -                    | 72.7            | -                    |
|              | VLP-MABSA (Ling, Yu, and Xia 2022)         | 78.6            | 73.8                 | 73.8            | 71.8                 |
| Text + Visual| Multi-BART (Lewis et al. 2020)             | 77.2            | 72.6                 | 70.5            | 69.0                 |
|              | SeqCSG (Ours)                               | 79.3            | 75.0                 | 74.6            | 73.2                 |

4.4 Ablation study

Component Analysis. We conduct the ablation study to validate the effectiveness of each component in our proposed approach on Twitter2015. The difference between BART and Multi-BART is that Multi-BART takes the caption information of the image as the input. Instead, SeqCSG optimizes the input structure of the model and converts the problem into a classification problem under a generation-based paradigm. We observe that our model exhibits a performance decay in the absence of any one of the component, i.e., prompt template, multi-modal KG, parameter frozen, or multi-modal visible matrix, demonstrating that all the modules are advantageous.

In addition, the improvement in experimental results also benefited from the that we perform a MABSC task in a prompt-guided generation-based paradigm. We can further improve the correlation between the target and other information by building prompt templates, which are naturally adapted to the sentiment classification task. By matching the extracted sequential cross-model semantic graph with the prompt template, we enabled the model to focus on specific visual features for different targets.

Table 3: Ablation Study on Twitter2015 dataset. In contrast to BART, Multi-BART uses caption information from the image. On the basis of SeqCSG-base, prompt, multi-modal KG, freeze and visible matrix were successively added.

| Method                | Acc  | Macro-F1 |
|-----------------------|------|----------|
| BART                  | 76.0 | 67.6     |
| Multi-BART            | 77.2 | 72.6     |
| SeqCSG-base           | 77.9 | 72.8     |
| w/ {Prompt}           | 78.1 | 73.6     |
| w/ {Prompt & Multi-KG}| 78.2 | 74.4     |
| w/ {Prompt & Multi-KG & freeze} | 78.9 | 74.7     |
| SeqCSG(Ours)          | 79.3 | 75.0     |
2) The quantity and quality of triples have a great impact on the performance of the model. It can be seen from the experimental results that the model performs best when the number of recalled triples is controlled to five triples of entities and between entities and five triples of entities and relevant subimages. We believe that when the number of triples is too small, the triples knowledge provided by the picture is limited. When there are too many triples, on the one hand, it will affect the speed of training, on the other hand, it will carry more noise. Both of these situations limit the performance of the model.

5 Conclusion

In this paper, we propose an multi-modal aspect-based sentiment classification (MABSC) method SeqCSG which designs a multi-modal visible matrix for an extracted semantic graph and transforms all elements of the semantic graph into a sequence format. Experimental results show that our proposed approach generally outperforms the state-of-the-art methods on standard benchmarks. As a unified model, SeqCSG integrates both the benefits of prompts and sequential cross-modal semantic graphs, which effectively focuses on specific image features for different targets.

In the future, we plan to (1) apply our approach to more image-enhanced natural language processing and information retrievals tasks, such as multi-modal event extraction and multi-modal entity retrieval; (2) incorporate the conceptual KG, and unify the image-level information and conceptual knowledge to perform joint reasoning of the scene, which is applied to tasks like visual question answering (Chen et al. 2021c); and (3) put attention to low resource scenarios (Chen et al. 2021a,b; Geng et al. 2021) with less or even no training data.

Figure 3: Predictions of Multi-BART, CapTrBERT, ours (w/o multi KG & visible matrix), ours (w/o visible matrix) and ours (SeqCSG) on several test samples. Specially, we provide a sequential cross-modal semantic graph for each sample.

Figure 4: Performance of different triple numbers on Twitter2015 dataset for MABSC task.

4.5 Error Analysis

Case Study. To further analyze the robustness of our method for error sensitivity, we visualize some predictions from different methods. The compared method contains BART, CapTrBERT, our model using the same inputs without multi-modal KG and visible matrix, and our model using multi-modal KG without visible matrix, respectively. As illustrated in Figure 3, BART outputs wrong predictions in all these four cases. CapTrBERT outputs correct prediction in the third case but makes mistakes in the first, the second and the fourth cases, where the caption can not provide enough information from images. In contrast, our full model, which combines the multi-modal KG and the visible matrix, makes correct predictions in those cases. Among all the cases, we notice that our model obtains more fine-grained multi-modal representation, which is essential for reducing error sensitivity. We can further reveal that the model lacking multi-modal KG and visible matrix has a poor prediction effect in some cases, which shows the superiority of our framework and semantic graph construction.
References

Chen, J.; Geng, Y.; Chen, Z.; Horrocks, I.; Pan, J. Z.; and Chen, H. 2021a. Knowledge-aware Zero-Shot Learning: Survey and Perspective. In IJCAI, 4366–4373. ijcai.org.

Chen, J.; Geng, Y.; Chen, Z.; Pan, J. Z.; He, Y.; Zhang, W.; Horrocks, I.; and Chen, H. 2021b. Low-resource Learning with Knowledge Graphs: A Comprehensive Survey. CoRR, abs/2112.10006.

Chen, Y.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: UNiversal Image-TEXT Representation Learning. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX, volume 12375 of Lecture Notes in Computer Science, 104–120. Springer.

Chen, Z.; Chen, J.; Geng, Y.; Pan, J. Z.; Yuan, Z.; and Chen, H. 2021c. Zero-Shot Visual Question Answering Using Knowledge Graph. In ISWC, volume 12922 of Lecture Notes in Computer Science, 146–162. Springer.

Chen, Z.; Huang, Y.; Chen, J.; Geng, Y.; Fang, Y.; Pan, J. Z.; Zhang, N.; and Zhang, W. 2022a. LaKo: Knowledge-driven Visual Question Answering via Late Knowledge-to-Text Injection. CoRR, abs/2207.12888.

Chen, Z.; Huang, Y.; Chen, J.; Geng, Y.; Zhang, W.; Fang, Y.; Pan, J. Z.; Song, W.; and Chen, H. 2022b. DUET: Cross-modal Semantic Grounding for Contrastive Zero-shot Learning. CoRR, abs/2207.01328.

Dai, J.; Yan, H.; Sun, T.; Liu, P.; and Qiu, X. 2021. Does syntax matter? A strong baseline for Aspect-based Sentiment Analysis with RoBERTa. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tür, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, 1816–1829. Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 4171–4186. Association for Computational Linguistics.

Fan, F.; Feng, Y.; and Zhao, D. 2018. Multi-grained Attention Network for Aspect-Level Sentiment Classification. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsuji, J., eds., Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, 3433–3442. Association for Computational Linguistics.

Geng, Y.; Chen, J.; Chen, Z.; Pan, J. Z.; Yuan, Z.; and Chen, H. 2021. K-ZSL: Resources for Knowledge-driven Zero-shot Learning. CoRR, abs/2106.15047.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 770–778. IEEE Computer Society.

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.; Huang, X.; Specia, L.; and Yih, S. W., eds., Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, 4395–4405. Association for Computational Linguistics.

Keswani, V.; Singh, S.; Agarwal, S.; and Modi, A. 2020. IITK at SemEval-2020 Task 8: Unimodal and Bimodal Sentiment Analysis of Internet Memes. In Herbelot, A.; Zhu, X.; Palmer, A.; Schneider, N.; May, J.; and Shutova, E., eds., Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020, 1135–1140. International Committee for Computational Linguistics.

Khan, Z.; and Fu, Y. 2021. Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Transformation. In Shen, H. T.; Zhuang, Y.; Smith, J. R.; Yang, Y.; Cesari, P.; Metze, F.; and Prabhakaran, B., eds., MM ’21: ACM Multimedia Conference, Virtual Event, China, October 20-24, 2021, 3034–3042. ACM.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In Meila, M.; and Zhang, T., eds., Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, 5583–5594. PMLR.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, 7871–7880. Association for Computational Linguistics.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX, volume 12375 of Lecture Notes in Computer Science, 121–137. Springer.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022, 2149–2159. Association for Computational Linguistics.
Appendix

A. Details of SeqCSG

A.1 Data Pre-processing

The input of the encoder consists of triple sentence, caption and tweet text. Caption is mainly to capture the global information of the image. Considering that the tweet text comes from Twitter, the text contains many special symbols, we perform data cleaning on it. Meanwhile, we replace the $T$ representing the target in the text with the specific target name, and add $\langle$target$\rangle$ and $\langle$/target$\rangle$ before and after the specific target in the text. The purpose is to guide the model to focus on target and target-related content. The triple sentence consists of serialized triples separated by a special token $\langle$ts$\rangle$. And each triple is converted into texts by connecting the head entity, relationship, and tail entity with a comma.

Considering that the above three parts together constitute the input of the encoder, we connect them in the following two ways,

$$\langle s \rangle \langle \text{triple} \rangle \text{triple sentence} \langle /\text{triple} \rangle \langle \text{caption} \rangle \text{caption} \langle /\text{caption} \rangle \langle \text{tweet} \rangle \text{tweet} \langle /\text{tweet} \rangle \langle /s \rangle$$

(9)

$$\langle s \rangle \text{triple sentence} \langle /s \rangle \text{caption} \langle /s \rangle \text{tweet} \langle /s \rangle$$

(10)

A.2 Multi-modal Visible Matrix Construction

In this section, we provide the details of the multi-modal visible matrix construction. We construct a relatedness matrix to indicate the relationship between relevant tokens for the input that contains serialized triples, tweets and captions on the encoder side.

We formulate different rule constraints for serialized triplet sentences, captions, tweet texts and special tokens. For the tweet text, caption and other special tokens, we require them to be visible, so that the text information can interact with the triplet information generated by the image. The visible matrix between triplet sentences is very critical because it can establish implicit relationships between targets while restoring the triplet structure. In addition, restricting the invisible relationship between some entities and relationship tokens can reduce the noise of the model. The specific multi-modal visible matrix construction is shown in Figure 5.

A.3 Algorithm Pseudo Code

Algorithm 1 describes the SeqCSG process. Our input contains three parts: an image $v$, a tweet text $s$ and an opinion target $t$. For the encoder end, we first apply the captioning transformer to turn the image $v$ into corresponding image caption $C$. Meanwhile, we parse a scene graph $G$ from the image $v$ and serialize these triples. Then we transform the semantic graph, which contains the triple sentence, the caption and the tweet text into a sequence. Multi-modal visible matrix is designed to construct the structure of the semantic graph. For the decoder end, we build a prompt template based on the target. Finally, we classify and predict the embedding output of $\langle$mask$\rangle$ position.
Figure 5: Visible matrix construction example to indicate the relationship between relevant tokens for the input that contains serialized triples, tweets and captions.

Algorithm 1: SeqCSG algorithm.

**Input:** The image \( v \); tweet text \( s \); opinion target \( t \); image caption function \( S_C \); scene graph parsing function \( S_{GP} \); serialize function \( Z \); visible matrix construction function \( V_{MC} \); prompt template construction function \( P_{TC} \); encoder decoder function \( ED \).

**Output:** Target’s classification \( \hat{y} \).

1: \( C \leftarrow S_C(v) \); \( G \leftarrow S_{GP}(v) \)
2: \( S \leftarrow Z(G) \)
3: \( S_{encoder} \leftarrow S + \langle/s\rangle + C + \langle/s\rangle + t \)
4: \( VM \leftarrow V_{MC}(S_{encoder}) \)
5: \( S_{decoder} \leftarrow P_{TC}(t) \)
6: \( h \leftarrow ED(S_{encoder}, S_{decoder}, VM) \)
7: \( \hat{y} \leftarrow MLP(h_{[n]} \)

B Details about Experimental Setup

B.1 Dataset Descriptions

We adopt two benchmark datasets annotated by [Yu and Jiang (2019)](#), namely Twitter2015 and Twitter2017 for MABSC to evaluate our model. Twitter2015 and Twitter2017 are two multi-modal datasets respectively collected by [Zhang et al. (2018)](#) and [Lu et al. (2018)](#). Both datasets are constructed similarly. Take Twitter2015 as an example, they use Twitter’s API to collect the tweets. The collection includes 26.5 million tweets. Then they drop the non-English tweets and extract containing relevant images from all those tweets, leaving 4.3 million tweets. They randomly sampled 50,000 data containing images from tweets covering various topics to reduce user-introduced specificity. Follow the standard annotation naturally, annotators annotate entities whose entity types are Person, Location, Organization, or Miscellaneous. On this basis, in order to solve the MABSC task, [Yu and Jiang (2019)](#) asks three domain experts to annotate the sentiment towards each target, and take the majority label among the three annotators as the gold label. Basic statistics have shown in Table 1.

B.2 Baselines

We adopt three types of baselines. The details of each baseline are listed below:

### Image Only Methods.

- **Res-Target** [He et al. (2016)](#) directly applies cross-modal attention to ResNet input features as the language features without any extra modifications.

### Text Only Methods.

- **MGAN** [Fan, Feng, and Zhao (2018)](#) proposes a fine-grained attention mechanism, which is responsible for linking and fusing the words from the aspect and context. Then this model combine it with the coarse-grained attention mechanism in order to capture the word-level interaction.
- **BERT** [Devlin et al. (2019)](#) is a simple baseline that only uses BERT encoder.
- **BERT+BL** [Devlin et al. (2019)](#) is BERT with another BERT layer stacked on it.
- **BERT-Pair-QA** [Sun, Huang, and Qiu (2019)](#) uses the auxiliary question method to obtain SOTA on SemEval 2014 Task 4.
- **BART** [Lewis et al. (2020)](#) is a baseline under a generation-based paradigm, which only takes text and target as input.

### Text and Image Methods.

- **Res-MGAN** [Fan, Feng, and Zhao (2018)](#) uses a multi-grain attention network for aspect understanding.
- **Res-BERT+BL** [Devlin et al. (2019)](#) directly applies cross-modal attention to ResNet input features and the language features without any extra modifications.
- **TomBERT** [Yu and Jiang (2019)](#) is also a target-oriented multi-modal BERT model. TomBERT builds on top of the baseline BERT architecture by adding target-sensitive visual attention and more self-attention layers to capture cross-modal dynamics.
- **CapTrBERT** [Khan and Fu (2021)](#) optimizes on TomBERT, but transforms image information into image caption, and then fuses the information of the two modes.
- **JML-MASC** [Ju et al. (2021)](#) is a multi-task learning approach proposed recently with the auxiliary cross-modal relation detection task.
- **VLP-MABSA** [Ling, Yu, and Xia (2022)](#) proposes a task-specific Vision Language Pre-training framework for MABSA.
- **Multi-BART** [Lewis et al. (2020)](#) is strong baseline under a generation-based paradigm. Unlike BART, Multi-BART uses the information of two modalities of image and text, and takes text, target and caption generated by image as the input of the model.
B.3 Downstream Details

This section contains details about training procedures and hyper-parameters for each dataset. We utilize Pytorch to conduct experiments with a RTX3090 GPU. All optimizations are performed with the AdamW optimizer with a linear warmup of learning rate.

Specifically, we use BART-base as our framework. The encoder and decoder both have six layers and are initialized with BART-base parameters. We employ ResNet to encoder image features and set ResNet’s parameters freeze, leaving only one linear layer to learn during training. We detail the hyper-parameter as follows:

**Twitter2015**
- max epoch: 30
- batch size: [8, 16]
- learning rate: [1e-5, 2e-5, 3e-5]
- image encoder: [ResNet18, ResNet34, ResNet50, ResNet101]
- number of triples between objects: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
- number of triples between object and image: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

**Twitter2017**
- max epoch: 30
- batch size: [8, 16]
- learning rate: [1e-5, 2e-5, 3e-5]
- image encoder: [ResNet18, ResNet34, ResNet50, ResNet101]
- number of triples between objects: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
- number of triples between object and image: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

C Additional Experimental Results

C.1 Effect of Image Encoder

We further analyze the influence of image encoder on Twitter2015. As shown in Table 4, we can draw the following conclusions: (1) Different image encoding modules have improved the effect of the model. (2) Different ResNet encoders have different effects on the experimental results. This is mainly because the granularity of the sub-images we extract from the picture is different, which leads to different effective information brought by different encoders.

| Method                  | Acc | Macro-F1 |
|-------------------------|-----|----------|
| w/o [Multi-KG & visible matrix] | 78.1 | 73.6     |
| SeqCSG(ResNet18)        | 78.7 | 75       |
| SeqCSG(ResNet34)        | 78.7 | 75.1     |
| SeqCSG(ResNet50)        | **79.3** | 75       |
| SeqCSG(ResNet101)       | 78.9 | **75.4** |

Table 4: Performance of different image encoders on Twitter2015 dataset for MABSC task.

C.2 Interpretability Analysis

Figure 6 visualizes the cross attention between \(<img>\) token in encoder and \(<mask>\) token in decoder maps on the case. Through the visualization of the case, we can notice the cross attention weights reveal that our model can capture the fine-grained semantics of the image. More importantly, our model can learn the implicit correlation representation of the target and the relevant subimage. We can draw the conclusion that irrelevant visual features may hurt the performance, while our model is able to benefit more fine-grained and implicit multi-modal representation, which is essential for reducing error sensitivity.

Figure 6: The cross attention visualizations between \(<img>\) token in encoder and \(<mask>\) token in decoder.