

Targeted Multimodal Sentiment Classification Based on Coarse-to-Fine Grained Image-Target Matching

Jianfei Yu*, Jieming Wang*, Rui Xia[†] and Junjie Li

School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{jfyu, wjm, rxia, jj_li}@njjust.edu.cn

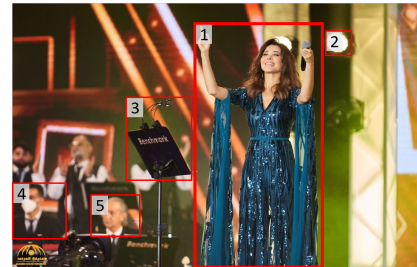
Abstract

Targeted Multimodal Sentiment Classification (TMSC) aims to identify the sentiment polarities over each target mentioned in a pair of sentence and image. Existing methods to TMSC failed to explicitly capture both coarse-grained and fine-grained image-target matching, including 1) the relevance between the image and the target and 2) the alignment between visual objects and the target. To tackle this issue, we propose a new multi-task learning architecture named coarse-to-fine grained Image-Target Matching network (ITM), which jointly performs image-target relevance classification, object-target alignment, and targeted sentiment classification. We further construct an Image-Target Matching dataset by manually annotating the image-target relevance and the visual object aligned with the input target. Experiments on two benchmark TMSC datasets show that our model consistently outperforms the baselines, achieves state-of-the-art results, and presents interpretable visualizations.¹

1 Introduction

As an important fine-grained task in multimodal sentiment analysis, Targeted Multimodal Sentiment Classification (TMSC, a.k.a aspect-based multimodal sentiment classification) has received increasing attention in recent years. Given a pair of sentence and image, the goal of TMSC is to identify the sentiment polarities towards each opinion target in the sentence [Xu *et al.*, 2019b; Yu *et al.*, 2019]. For example, in Fig. 1, given the multimodal tweet and its two opinion targets “Nancy Ajram” and “Salalah Tourism Festival”, it is expected to identify that the user expresses *Positive* and *Neutral* sentiments towards them, respectively.

In the literature, a myriad of deep learning approaches have been proposed for the TMSC task. [Xu *et al.*, 2019b] and [Yu *et al.*, 2020] focused on designing effective attention mechanisms to model the interactions among the target, text, and



[Nancy Ajram]_{Positive, Box-1} during the [Salalah Tourism Festival]_{Neutral, Box-N/A}; beautiful as always.

Figure 1: An Example of Targeted Multimodal Sentiment Classification (TMSC). *Nancy Ajram* and *Salalah Tourism Festival* are two mentioned targets. Box-1 denotes the 1st bounding box is aligned with *Nancy Ajram*, and Box-N/A denotes no bounding box is aligned with *Salalah Tourism Festival*.

image. [Yu *et al.*, 2019] and [Wang *et al.*, 2021] followed the recent pre-train and fine-tune paradigm, and adapted existing pre-trained models to capture the text-image, target-text, and target-image interactions. More recently, [Khan *et al.*, 2021] proposed a Transformer-based image captioning model to translate the image to an auxiliary sentence, and then combined the original and auxiliary sentences for targeted sentiment classification.

However, all these existing studies failed to explicitly consider the matching relations between the target and the image, which is essential for the TMSC task for following reasons:

- **Coarse-Grained Image-Target Matching.** Based on our observations of a benchmark Twitter dataset of TMSC, around 58% of the input targets are not presented in associated images in a benchmark dataset, and these unrelated images will inevitably bring much noise for the TMSC task. For example, in Fig. 1, given *Salalah Tourism Festival* as the input target, the unrelated image may mislead the model to predict its sentiment as *Positive*. Hence, it is crucial to capture the image-target relevance to alleviate the visual noise for targeted sentiment classification.
- **Fine-Grained Image-Target Matching.** For those target-related images, as each image contains a number of visual objects (i.e., fine-grained image), identifying the aligned visual object to the input target is generally helpful for pre-

*Equal contribution.

[†]Corresponding author.

¹The source code is released at <https://github.com/NUSTM/ITM>.

dicting its sentiment. For example, in Fig. 1, among all the marked bounding boxes, the bounding box with the pleasant woman (i.e., Box-1) provides the most important clue for detecting the *Positive* sentiment over *Nancy Ajram*.

Motivated by these observations, we propose a coarse-to-fine grained Image-Target Matching network (ITM) for the TMSC task. Specifically, we first construct an Image-Target Matching dataset by manually annotating 1) the relevance between the image and the target and 2) the visual object (i.e., bounding box) aligned with the input target. With such an annotated dataset, we propose a multi-task learning architecture ITM to jointly perform coarse-to-fine grained image-target matching and targeted sentiment classification. ITM contains three key modules: the first module is to identify image-target relevance for dynamically controlling the contribution of visual information; with the filtered visual information, the second module focuses on object-target alignment to learn appropriate weights of each visual object based on their alignment probabilities with the input target; the last module performs multimodal fusion and sentiment classification.

Experimental results on two benchmark datasets for the TMSC task show that our multi-task learning model ITM consistently outperforms a number of state-of-the-art methods, and presents insightful and interpretable visualizations, demonstrating the importance of coarse-grained and fine-grained image-target matching to the TMSC task.

2 Task Formulation

Given a multimodal corpus \mathbb{D} , let us first use $\{X_1, X_2, \dots, X_{|\mathbb{D}|}\}$ to denote a set of samples in the corpus. For each sample, we are given an n -word sentence $\mathbf{S} = (w_1, w_2, \dots, w_n)$, an image \mathbf{V} , and an m -word opinion target $\mathbf{T} = (t_1, t_2, \dots, t_m)$, where \mathbf{T} is a sub-sequence of \mathbf{S} . We then formulate the three tasks in our work as follows:

Image-Target Relevance. For each sample $X = (\mathbf{S}, \mathbf{V}, \mathbf{T})$, the target \mathbf{T} is assumed to be associated with a relevance label r indicating whether the image \mathbf{V} is related to \mathbf{T} , where r is either *Related* or *Unrelated*. The goal of this task is to learn a binary classification function that maps X to r .

Object-Target Alignment. For each sample $X = (\mathbf{S}, \mathbf{V}, \mathbf{T})$, an object detection method is employed to identify K object proposals in the image \mathbf{V} , and the target \mathbf{T} is associated with its alignment distribution over the K object proposals, denoted by \mathbf{A} . The goal of this task is to learn a mapping from X to the alignment distribution \mathbf{A} .

TMSC. For each sample $X = (\mathbf{S}, \mathbf{V}, \mathbf{T})$, we assume that the target \mathbf{T} is associated with a sentiment label y , which can be *Positive*, *Negative* or *Neutral*. The goal of this main task is to learn a sentiment classifier that maps X to y .

3 Dataset

We construct an Image-Target Matching dataset for Image-Target Relevance and Object-Target Alignment tasks.

Source. Since both tasks require the annotation of targets, we construct our dataset based on a subset of one benchmark dataset for the TMSC task (i.e., TWITTER-17), which has

Split	#Targets	#Images	#I-T Related	#I-T Unrelated	#Annotated Boxes
Train	1176	600	459	717	459
Dev	588	297	254	334	254
Test	588	280	270	318	270
Total	2352	1177	983	1369	983

Table 1: Statistic of Our Image-Target Matching Dataset.

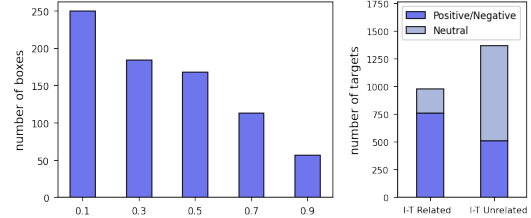


Figure 2: The Box/image area ratio (left) and the correlation of Image-Target (I-T) relevance and sentiment (right) in our dataset.

annotated the targets by [Lu *et al.*, 2018]. We randomly select 1176, 588, and 588 samples from the training, development, and test sets of TWITTER-17, and employ two PhD students for annotation. The annotation for the Image-Target Relevance task reaches an agreement of 98.5%, and the agreement for bounding box annotation is 92.3%, indicating the high quality of our data. For disagreement samples, we ask a third expert to make the final decision.

Statistics and Analysis. The basic statistic of our dataset is shown in Table 1. It can be seen that a large percentage of targets are unrelated to images. For each target-related image, since the semantic meaning of the target is clear, only one bounding box is annotated. Fig. 2 (left) shows the distribution of bounding box area over image area ratio. Compared to images, most bounding boxes are relatively small, which implies the challenge of object-target alignment. In Fig. 2 (right), we further show the correlation between sentiment and image-target relevance. It is interesting to observe that for targets related to the images, users tend to express either positive or negative sentiment towards them; whereas for targets unrelated to the images, users tend to express neutral sentiment over them. This indicates image-target relevance indeed provides important clues to TMSC.

4 Methodology

We propose a multi-task learning framework named coarse-to-fine grained Image-Target Matching network (ITM), which leverages two auxiliary tasks, i.e., image-target relevance and object-target alignment, to improve the TMSC task. As shown in Fig. 3, ITM consists of four modules: Feature Extraction, Coarse-Grained Matching, Fine-Grained Matching, and Multimodal Fusion. We describe the details of each module in the following subsections.

4.1 Feature Extraction

Contextualized Target Representation. Given an input sentence \mathbf{S} and its target \mathbf{T} , we split \mathbf{S} into two parts, i.e., the target and the remaining context, and combine them as

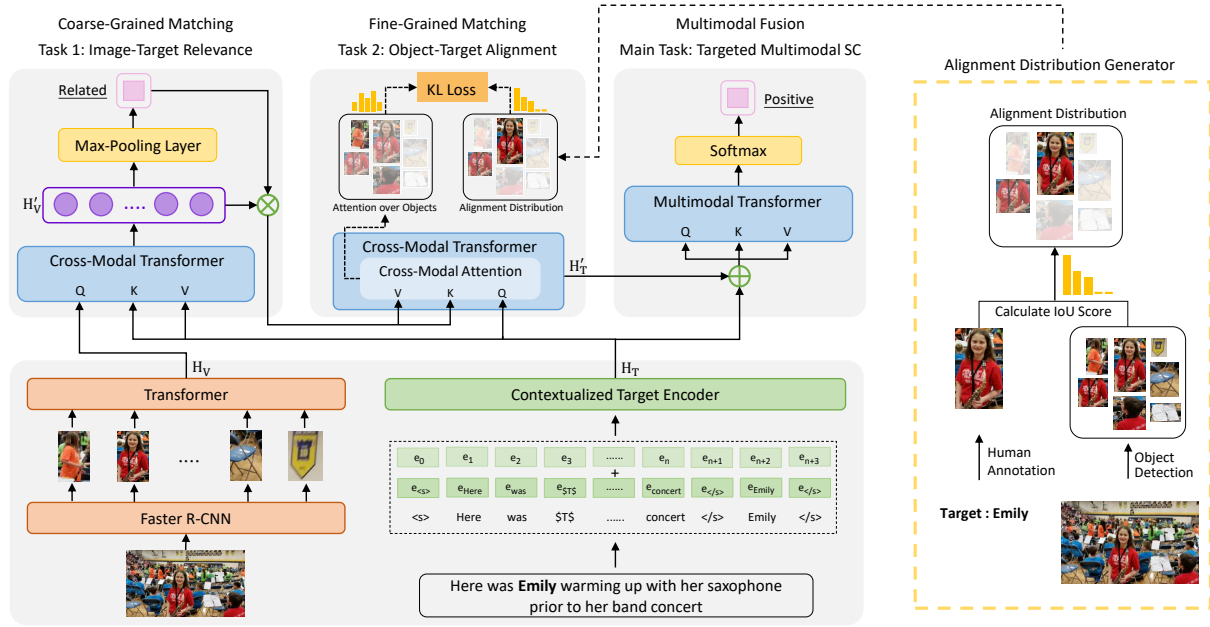


Figure 3: (a) Overview of Coarse-to-Fine Grained Image-Target Matching Network.

(b) Alignment Distribution.

the contextualized target input T' . An example is shown in the bottom of Fig. 3.a, where we replace the target *Emily* in S with a special token $\$T\$$ as its context, and concatenate the context with *Emily* as the input. With the transformed target input T' , we feed it to a widely-used pre-trained model RoBERTa [Liu *et al.*, 2019] to obtain the contextualized target representation: $H_T = \text{RoBERTa}(T')$, where $H_T \in \mathbb{R}^{d \times n}$, d is the hidden dimension, and n is the input length.

Image Representation. Given an image V , we use a widely-used object detection method Faster R-CNN [Ren *et al.*, 2015] to detect object proposals and obtain their regional representations as our visual features [Anderson *et al.*, 2018]. We then sort detected proposals by object category detection probabilities, and keep the top-100 object proposals in order to retain more small objects for target alignment. Let $R = \text{Faster R-CNN}(V)$ denote the regional representations, where $R \in \mathbb{R}^{2048 \times 100}$. To model the interactions between objects, we feed R to Transformer to obtain object-level image representations: $H_V = \text{Transformer}(W_R^T R)$, where $W_R \in \mathbb{R}^{2048 \times d}$ and $H_V \in \mathbb{R}^{d \times 100}$.

4.2 Coarse-Grained Matching

The goal of this module is to capture the image-target relevance, and alleviate the noise from unrelated images.

To achieve this goal, we apply the Cross-Modal Transformer layer [Tsai *et al.*, 2019] to model the interaction between the target and the image, which regards image representations H_V as queries, and contextualized target representations H_T as keys and values as follows:

$$H'_V = \text{CM-Transformer}(H_V, H_T, H_T), \quad (1)$$

where $H'_V \in \mathbb{R}^{d \times 100}$ is the generated target-based image representation. Next, we apply a max-pooling operator over H'_V

to obtain the most salient features for relevance classification: $h'_V = \max\text{-pooling}(H'_V)$. Based on h'_V , we use a Sigmoid function to perform image-target relevance classification:

$$P(r) = \text{Sigmoid}(W_r h'_V + b_r). \quad (2)$$

We use the cross-entropy loss to optimize the image-target relevance task, denoted by Relevance (RE) Supervision:

$$\mathcal{L}^{RE} = -\frac{1}{M} \sum_{k=1}^M \log P(r^k), \quad (3)$$

where M is the number of samples in our annotated dataset.

Since the probability in Eqn. (2) is a scalar in the range of $[0,1]$ indicating the relevant score between the target and the image, we use it to construct a visual filter matrix $G \in \mathbb{R}^{d \times 100}$, where each entry in G equals to $P(r)$. With the visual filter matrix, we can obtain the filtered image representations as follows:

$$H''_V = G \odot H'_V. \quad (4)$$

where \odot is the element-wise multiplication. For example, if G equals to $\mathbf{0}$, all the visual features are filtered.

4.3 Fine-Grained Matching

Based on coarse-grained image-target matching, this Fine-Grained Matching module further aims to identify the fine-grained visual objects aligned with the input target in those target-related images.

To achieve the object-target alignment, we apply another Cross-Modal Transformer layer to obtain the target-aware attention distribution over 100 object proposals from Faster R-CNN. Specifically, we use the representation of the first token in the target input (i.e., H_T^0) as queries, and the filtered image representations H''_V as keys and values:

$$H'_T = \text{CM-Transformer}(H_T^0, H''_V, H''_V), \quad (5)$$

where $\mathbf{H}'_{\mathbf{T}} \in \mathbb{R}^{d \times 1}$ is the generated image-based target representations. Let us use \mathbf{D}_i to denote the attention weights in the i -th head attention of the Cross-Modal Transformer. We take the average of all the m -head attentions as the final distribution over 100 object proposals, denoted by $\mathbf{D} = \frac{1}{m} \sum_{i=1}^m \mathbf{D}_i$, where $\mathbf{D} \in \mathbb{R}^{100}$.

To guide the attention distribution to achieve object-target alignment, we propose to obtain an alignment distribution from ground-truth (GT) boxes as supervision. As shown in Fig 3.b, given an image, we first calculate the Intersection over Union (IoU) scores of its object proposals with respect to the GT bounding box, which denote the overlap between the proposal and GT bounding box. Following previous studies for visual grounding [Yu *et al.*, 2018; Lei *et al.*, 2020], for the i -th proposal, if its IoU score is larger than 0.5, we keep the IoU score and 0 otherwise. We can then get the IoU score distribution over all object proposals, denoted by $[a_1, \dots, a_{100}] \in \mathbb{R}^{100}$. Based on this, we re-normalize the IoU score distribution to obtain the GT alignment distribution $\mathbf{A} \in \mathbb{R}^{100}$.

We adopt the Kullback-Leibler Divergence (KLD) loss to make the attention distribution \mathbf{D} and the ground-truth alignment distribution \mathbf{A} as close as possible, denoted by Attention (ATT) Supervision:

$$\mathcal{L}^{ATT} = \frac{1}{C} \sum_{i=1}^C \mathbf{A}^i \log\left(\frac{\mathbf{A}^i}{\mathbf{D}^i}\right). \quad (6)$$

where C is the number of target-image related samples in our Image-Target Matching dataset.

4.4 Multimodal Fusion

With the image-based target representations $\mathbf{H}'_{\mathbf{T}}$ generated from the Fine-Grained Matching module, we concatenate it with the contextualized target representations as: $\mathbf{H}_{\mathbf{M}} = \mathbf{H}'_{\mathbf{T}} \oplus \mathbf{H}_{\mathbf{T}}$, and feed them to a Transformer layer for multimodal fusion:

$$\mathbf{H} = \text{MM-Transformer}(\mathbf{H}_{\mathbf{M}}, \mathbf{H}_{\mathbf{M}}, \mathbf{H}_{\mathbf{M}}), \quad (7)$$

Finally, the representation of the first token is fed to a softmax layer for sentiment classification:

$$P(y) = \text{Softmax}(\mathbf{W}^T \mathbf{H}^0 + \mathbf{b}). \quad (8)$$

The standard cross-entropy loss is to optimize the TMSC task, denoted by Sentiment Supervision:

$$\mathcal{L}^{TMSC} = -\frac{1}{N} \sum_{j=1}^N \log P(y^j) \quad (9)$$

where N is the number of samples for the TMSC task.

We employ the alternating optimization strategy to iteratively optimize the two auxiliary tasks with our Image-Target Matching dataset and optimize the main task with the dataset for TMSC. The combined objective function is:

$$\mathcal{J} = \lambda_1 \mathcal{L}^{RE} + \lambda_2 \mathcal{L}^{ATT} + \mathcal{L}^{TMSC} \quad (10)$$

where λ_1 and λ_2 are hyper-parameters.

Label	TWITTER-15			TWITTER-17		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	517	573
Negative	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234

Table 2: Statistics of two benchmark datasets for TMSC.

5 Experiment

5.1 Experiment Setting

We adopt three datasets to systematically evaluate the effectiveness of our coarse-to-fine grained Image-Target Matching network (ITM). One is our Image-Target Matching dataset for the two auxiliary tasks, i.e., Image-Target Relevance and Object-Target Alignment, as introduced in Section 3. The other two are the benchmark Twitter datasets for the TMSC task, i.e., TWITTER-15 and TWITTER-17. The statistics of the two TMSC datasets are shown in Table 2.

For our ITM model, we adopt RoBERTa_{base} [Liu *et al.*, 2019] as the contextualized target encoder and Faster R-CNN [Ren *et al.*, 2015] with ResNet-101 backbone released by [Anderson *et al.*, 2018] as the object detector. During the alternating optimization process, we use the AdamW optimizer, and fix the hyper-parameters after tuning them on the development set. Specifically, we set the batch size to 32, the training epoch to 10, and λ_1 and λ_2 to 1 and 0.5. The learning rates for the TMSC task and the two auxiliary tasks are set to 1e-5 and 1e-6 respectively.

5.2 Main Results

In this subsection, we compare our ITM model with several representative methods for TMSC, and report the accuracy (Acc) and the Macro-F1 score (F1) of each method in Table 3.

We first consider the following methods that focus on text only for comparison: 1) *MGAN* [Fan *et al.*, 2018], a multi-grained attention network capturing multi-level target-text interactions. 2) *BERT* [Devlin *et al.*, 2019], a pre-trained model regarding target and text as a pair for sentiment classification. 3) *RoBERTa* [Liu *et al.*, 2019], an enhanced pre-trained model based on BERT. Moreover, we consider the following multimodal approaches for comparison: 1) *MIMN* [Xu *et al.*, 2019b], a multi-interactive memory network modeling the interaction between the target, text and image. 2) *ESAFN* [Yu *et al.*, 2020], a target-sensitive attention and fusion network based on LSTM. 3) *ViLBERT* [Lu *et al.*, 2019], a pre-trained Vision-Language model, in which the target-text pair is used as the textual input. 4) *CapBERT* [Khan *et al.*, 2021], which translates the image to textual captions and combines the captions and the original target-text pair with a pre-trained BERT model. 5) *TomBERT* [Yu *et al.*, 2019], a BERT-based TMSC approach with target-sensitive cross-modal attention. 6) *CapRoBERTa*, which replaces BERT with RoBERTa in *CapBERT*. 7) *TomRoBERTa*, which replaces BERT and ResNet with RoBERTa and Faster R-CNN in *TomBERT*. 8) *TomRoBERTa+Aux-Tasks*, a *TomRoBERTa*-based multi-task learning baseline proposed by us, which

Methods	TWITTER-15		TWITTER-17	
	Acc	F1	Acc	F1
Text Only				
MGAN [Fan <i>et al.</i> , 2018]	71.17	64.21	64.75	61.46
BERT [Devlin <i>et al.</i> , 2019]	74.15	68.86	68.15	65.23
RoBERTa [Liu <i>et al.</i> , 2019]	76.28	71.36	69.77	68.00
Text and Image				
MIMN [Xu <i>et al.</i> , 2019b]	71.84	65.69	65.88	62.99
ESAFN [Yu <i>et al.</i> , 2020]	73.38	67.37	67.83	64.22
ViLBERT [Lu <i>et al.</i> , 2019]	73.76	69.85	67.42	64.87
CapBERT [Khan <i>et al.</i> , 2021]	78.01±0.34	73.25±0.36	69.77±0.16	68.42±0.48
TomBERT [Yu <i>et al.</i> , 2019]	76.60±0.40	71.57±0.16	69.42±0.73	67.70±0.50
CapRoBERTa	77.82±0.43	73.38±0.48	71.07±0.49	68.57±0.55
TomRoBERTa	77.64±0.23	73.24±0.37	71.34±0.40	70.14±0.41
TomRoBERTa+Aux-Tasks	77.37±0.36	73.00±0.35	71.18±0.37	69.86±0.32
ITM (ours)	78.27±0.28	74.19±0.40	72.61±0.21	71.97±0.27

Table 3: Comparison between previous methods and our ITM model on two benchmark TMSC datasets. For the last 5 rows, we report the average results across three runs. \pm refers to standard deviations.

adds our attention supervision in Eqn. (6) on their target attention layer and adds a softmax layer with relevance supervision over their final multimodal representation.

In Table 3, we can observe that *RoBERTa* achieves the best performance among text-only methods. It is reasonable since *RoBERTa* adopted better training strategies and larger corpus than *BERT*. For multimodal methods, it is easy to see that *MIMN* and *ESAFN* obtain the lowest performance, due to the lack of model pre-training. The pre-trained VL model (i.e., *ViLBERT*) performs worse than *TomBERT*, probably because the pre-trained dataset for *ViLBERT* is much smaller than *BERT*. Moreover, *CapBERT* performs better than all the other baseline systems, since it resorts to a pre-trained image captioning model. It is intuitive that *TomRoBERTa* and *CapRoBERTa* generally performs better than *TomBERT* and *CapBERT*. In addition, it is surprising that *TomRoBERTa+Aux-Tasks* performs even worse than *TomRoBERTa*. We conjecture the reason is: 1) its target attention layer only uses the target without its context as the target input; 2) due to the structure of *TomBERT*, the object-target alignment is performed before the image-target relevance, which may bring much visual noise to object-target alignment. Finally, we can clearly see that ITM achieves the best results on both accuracy and F1 score among all the compared systems across the two datasets. These observations demonstrate the effectiveness of our ITM model and the importance of incorporating image-target matching for the TMSC task.

5.3 Results of Image-Target Matching

Table 4 shows the results of Image-Target Relevance Classification on our Image-Target Matching dataset in Section 3. It can be seen that our ITM model significantly outperforms *TomRoBERTa+Aux-Tasks* on all the metrics, showing the advantage of ITM for Image-Target Relevance.

Table 5 shows the results of Object-Target Alignment on our Image-Target Matching dataset. The evaluation metrics are Kullback-Leibler Divergence (KLD) between the attention distribution **D** and the ground-truth alignment distribution **A** in Section 4.3 and the recall of the top-ranked bound-

Methods	Acc	Precision	Recall	F1
TomRoBERTa+Aux-Tasks	71.09	68.70	68.04	68.36
ITM (ours)	71.94	71.75	71.76	71.76

Table 4: Performance on the Image-Target Relevance task.

Methods	KL Divergence	R@1	R@3	R@5
TomRoBERTa+Aux-Tasks	7.18	22.59	41.48	50.74
ITM (ours)	2.08	51.82	68.89	75.93

Table 5: Performance on the Object-Target Alignment task.

Methods	TWITTER-15		TWITTER-17	
	Acc	F1	Acc	F1
ITM	78.27	74.19	72.61	71.97
w/o Relevance (RE) Supervision	77.72	72.07	71.47	70.29
w/o Attention (ATT) Supervision	77.82	72.54	71.56	70.31
w/o RE & ATT Supervision	76.66	71.62	71.07	68.96
w/o Coarse-Grained Matching	76.95	71.70	71.39	70.16
w/o Fine-Grained Matching	77.14	71.65	71.34	70.35
w/o Coarse and Fine-Grained Matching	76.57	71.03	70.66	69.56

Table 6: Ablation study of our proposed model ITM.

ing box in **A** from top-*k* bounding boxes in **D**, denoted by $R@k$. In Table 5, it is clear that ITM significantly outperforms *TomRoBERTa+Aux-Tasks* in terms of all the metrics, showing the advantage of ITM for Object-Target Alignment.

5.4 In-depth Analysis

Ablation Study. We explore the impact of different components in our model and report the results in Table 6. Specifically, removing either the relevance supervision in Eqn. (3) or the attention supervision in Eqn. (6) leads to a moderate performance drop on both accuracy and F1 score. Moreover, discarding the two supervisions will lead to a significant performance drop of around 1.6 percentage points on accuracy and 3 percentage points on F1 score. These observations indicate the indispensable effects of filtering the visual noise and achieving object-target alignment. Lastly, from the last three rows of Table 6, we find that removing either coarse or fine-grained matching module or both modules in Section 4.2 and Section 4.3 consistently decreases the performance, which indicates the necessity of incorporating Cross-Modal Transformer layers to achieve cross-modal alignments.

Case Study. In the left two columns of Table 7, we show two representative test samples to demonstrate the importance of filtering the unrelated images. For case (a), given the target *Pacific Rim*, *RoBERTa* accurately predicted its sentiment as *Positive*, while *TomRoBERTa* gave the wrong prediction after combining the unrelated image. In contrast, ITM gave the correct sentiment prediction and a low image-target relevance score as well as an evenly distributed alignment distribution (i.e. no object is obviously aligned with *Pacific Rim* in the image). Similarly, for case (b), given *Stagecoach* as the target, *TomRoBERTa* wrongly predicted its sentiment as *Positive* due to the unrelated image, while ITM correctly predicted the *Neutral* sentiment after filtering the visual noise.


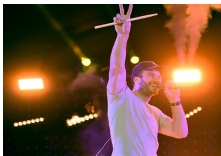



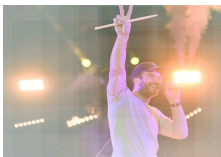
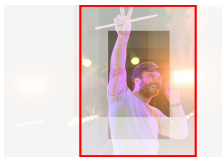
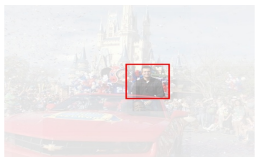
	Effects of Image-Target Relevance		Effects of Object-Target Alignment	
Image	I-T relevance score: 0.3478 	I-T relevance score: 0.2617 	I-T relevance score: 0.7524 	I-T relevance score: 0.7357 
Attention Weights				
Text	(a). Morning coffee before I go watch [Pacific Rim] _{Positive} # excited ...	(b). # SamHunt Performs at [Stagecoach] _{Neutral} # MusicFestival 2016	(c). # [SamHunt] _{Positive} Performs at Stagecoach # MusicFestival 2016	(d). RT @ DisneySports: [Joe Flacco] _{Positive} and Mickey Mouse at Magic Kingdomparade
RoBERTa	Positive ✓	Neutral ✓	Neutral ×	Neutral ×
TomRoBERTa	Neutral ×	Positive ×	Positive ✓	Neutral ×
ITM (ours)	Positive ✓	Neutral ✓	Positive ✓	Positive ✓

Table 7: Prediction comparison between different methods on four test samples. In the first row, we show the Image-Target related probability predicted by our ITM model. For the two Image-Target related samples in the right, the ground-truth (green), the top-ranked predicted object proposal (red) are visualized respectively. The second row visualizes the attention weights in the Fine-Grained Matching module of ITM.

In the right two columns of Table 7, we use another two test samples to show the effect of achieving object-target alignment. For case (c), we can see that it is in the same tweet in case (b), but the given target is changed to *SamHunt*. In this case, RoBERTa and TomRoBERTa still made the same predictions as before, while ITM generated the appropriate relevance score and assigned higher attention weights to the objects around *SamHunt*, and thus gave the correct sentiment. For case (d), given the target *Joe Flacco*, RoBERTa wrongly predicted its sentiment as *Neutral* due to the absence of sentiment words. TomRoBERTa also gave the wrong prediction, because it failed to identify the small object about Joe Flacco. By contrast, our ITM model correctly predicted the sentiment as *Positive* with more attention on his smiling face.

6 Related Work

Targeted sentiment classification (TSC, a.k.a aspect-based sentiment classification) has been well studied in recent years. Various traditional feature-based models [Jiang *et al.*, 2011; Pontiki *et al.*, 2016] and deep learning-based models [Tang *et al.*, 2016; Wang *et al.*, 2018; Xu *et al.*, 2020] have been proposed to address the TSC task. More recently, many Transformer-based methods [Xu *et al.*, 2019a; Dai *et al.*, 2021] and graph neural network-based methods [Wang *et al.*, 2020; Tang *et al.*, 2020] are designed to better leverage sequential and syntactic information for the task. Despite obtaining remarkable results, these approaches failed to consider the information from other modalities, e.g., images.

With the explosive growth of multimodal data, multimodal sentiment analysis (MSA) has attracted wide attention recently. For coarse-grained MSA, many approaches have explored the capability of adopting neural networks to build the interactions between modalities for MSA in conversations or tweets [Poria *et al.*, 2017; Xu *et al.*, 2018;

Zhang *et al.*, 2020; Yang *et al.*, 2021]. For fine-grained MSA, various LSTM-based and Transformer-based methods were proposed to capture the fine-grained interaction across different modalities for the TMSC task [Xu *et al.*, 2019b; Yu *et al.*, 2020]. In this work, we follow the later line of work, aiming to improve TMSC with Image-Target Matching.

Since our Object-Target Alignment task is closely related to Visual Grounding (VG), we review some representative studies for VG, which aim to predict the location of an image region referred by the language expression. Earlier works to VG primarily focus on selecting visual objects based on parsing linguistic descriptions [Kazemzadeh *et al.*, 2014; Yu *et al.*, 2016]. Recently, a myriad of visual-language pre-training models have been proposed for VG to capture the alignment between image and language modalities [Su *et al.*, 2019; Yu *et al.*, 2021].

7 Conclusion

In this paper, we proposed a multi-task learning model named coarse-to-fine grained Image-Target Matching network (ITM), which leverages two auxiliary tasks, i.e., Image-Target Relevance and Object-Target Alignment, to capture the image-target matching relations for the TMSC task. Experiment results on two TMSC datasets and our Image-Target Matching dataset demonstrate that our ITM model consistently outperforms a number of state-of-the-art methods.

Acknowledgments

This work was supported by the Natural Science Foundation of Jiangsu Province for Young Scholars (BK20200463) and Distinguished Young Scholars (BK20200018), and the Natural Science Foundation of China (62076133 and 62006117).

References

- [Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
- [Dai *et al.*, 2021] Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. Does syntax matter? a strong baseline for aspect-based sentiment analysis with roberta. In *NAACL*, 2021.
- [Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.
- [Fan *et al.*, 2018] Feifan Fan, Yansong Feng, and Dongyan Zhao. Multi-grained attention network for aspect-level sentiment classification. In *EMNLP*, 2018.
- [Jiang *et al.*, 2011] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *ACL*, 2011.
- [Kazemzadeh *et al.*, 2014] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, et al. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.
- [Khan *et al.*, 2021] Zaid Khan, Yun Fu, et al. Exploiting bert for multimodal target sentimentclassification through input space translation. In *ACM MM*, 2021.
- [Lei *et al.*, 2020] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *ACL*, 2020.
- [Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [Lu *et al.*, 2018] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *ACL*, 2018.
- [Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NIPS*, 2019.
- [Pontiki *et al.*, 2016] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *SemEval-2016*, 2016.
- [Poria *et al.*, 2017] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *ACL*, 2017.
- [Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.
- [Su *et al.*, 2019] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019.
- [Tang *et al.*, 2016] Duyu Tang, Bing Qin, Ting Liu, et al. Aspect level sentiment classification with deep memory network. In *EMNLP*, 2016.
- [Tang *et al.*, 2020] Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *ACL*, 2020.
- [Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 2019.
- [Wang *et al.*, 2018] Shuai Wang, Sahisnu Mazumder, Bing Liu, Mianwei Zhou, et al. Target-sensitive memory networks for aspect sentiment classification. In *ACL*, 2018.
- [Wang *et al.*, 2020] Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, et al. Relational graph attention network for aspect-based sentiment analysis. In *ACL*, 2020.
- [Wang *et al.*, 2021] Jiawei Wang, Zhe Liu, Victor Sheng, Yuqing Song, and Chenjian Qiu. Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification. In *PRCV*, 2021.
- [Xu *et al.*, 2018] Nan Xu, Wenji Mao, and Guandan Chen. A co-memory network for multimodal sentiment analysis. In *SIGIR*, 2018.
- [Xu *et al.*, 2019a] Hu Xu, Bing Liu, Lei Shu, et al. Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *NAACL*, 2019.
- [Xu *et al.*, 2019b] Nan Xu, Wenji Mao, and Guandan Chen. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *AAAI*, 2019.
- [Xu *et al.*, 2020] Lu Xu, Lidong Bing, Wei Lu, and Fei Huang. Aspect sentiment classification with aspect-specific opinion spans. In *EMNLP*, 2020.
- [Yang *et al.*, 2021] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. Multimodal sentiment detection based on multi-channel graph neural networks. In *ACL*, 2021.
- [Yu *et al.*, 2016] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.
- [Yu *et al.*, 2018] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. In *IJCAI*, 2018.
- [Yu *et al.*, 2019] Jianfei Yu, Jing Jiang, et al. Adapting bert for target-oriented multimodal sentiment classification. In *IJCAI*, 2019.
- [Yu *et al.*, 2020] Jianfei Yu, Jing Jiang, and Rui Xia. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM TASLP*, 28:429–439, 2020.
- [Yu *et al.*, 2021] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *AAAI*, 2021.
- [Zhang *et al.*, 2020] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Multimodal multi-label emotion detection with modality and label dependence. In *EMNLP*, 2020.