Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis

Yan Ling, Jianfei Yu*, and Rui Xia*
School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{ylin, jfyu, rxia}@njust.edu.cn

Abstract

As an important task in sentiment analysis, Multimodal Aspect-Based Sentiment Analysis (MABSA) has attracted increasing attention in recent years. However, previous approaches either (i) use separately pre-trained visual and textual models, which ignore the cross-modal alignment or (ii) use vision-language models pre-trained with general pre-training tasks, which are inadequate to identify fine-grained aspects, opinions, and their alignments across modalities. To tackle these limitations, we propose a task-specific Vision-Language Pre-training framework for MABSA (VLP-MABSA), which is a unified multimodal encoder-decoder architecture for all the pre-training and downstream tasks. We further design three types of task-specific pre-training tasks from the language, vision, and multi-modal modalities, respectively. Experimental results show that our approach generally outperforms the state-of-the-art approaches on three MABSA subtasks. Further analysis demonstrates the effectiveness of each pre-training task. The source code is publicly released at https://github.com/NUSTM/VLP-MABSA.

1 Introduction

Recent years have witnessed increasing attention on the Multimodal Aspect-Based Sentiment Analysis (MABSA) task[1]. Previous research mostly focused on its two subtasks, including Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect-oriented Sentiment Classification (MASC). Given a text-image pair as input, MATE aims to extract all the aspect terms mentioned in the text (Zhang et al., 2018; Lu et al., 2018; Wu et al., 2020a,b; Zhang et al., 2021a), whereas MASC aims to classify the sentiment towards each extracted aspect term (Xu et al., 2019; Yu and Jiang, 2019; Khan and Fu, 2021). As the two subtasks are closely related to each other, Ju et al. (2021) recently introduced the Joint Multimodal Aspect-Sentiment Analysis (JMASA) task, aiming to jointly extract the aspect terms and their corresponding sentiments. For example, given the text-image pair in Table. 1, the goal of JMASA is to identify all the aspect-sentiment pairs, i.e., (*Sergio Ramos*, *Positive*) and (*UCL*, *Neutral*).

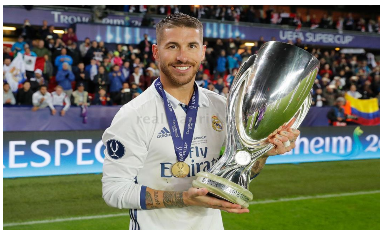
Image	
Text	Sergio Ramos chosen as the best player of UCL final
Output	(Sergio Ramos, Positive) (UCL, Neutral)

Table 1: An example of the MABSA task

Most of the aforementioned studies to MABSA primarily focused on employing pre-trained unimodal models (e.g., BERT for text and ResNet for image) to obtain textual and visual features respectively. The separate pre-training of visual and textual features ignores the alignment between text and image. It is therefore crucial to perform vision-language pre-training to capture such cross-modal alignment. However, for the MABSA task, the studies on vision-language pre-training are still lacking.

To the best of our knowledge, there are very few studies focusing on vision-language pre-training for one of the MABSA subtasks, i.e., MATE (Sun et al., 2020, 2021). One major drawback of these studies is that they mainly employ general vision-language understanding tasks (e.g., text-image

[*] Corresponding authors.

[1] The MABSA task is also known as Target-Oriented Multimodal Sentiment Analysis or Entity-Based Multimodal Sentiment Analysis in the literature.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2149 - 2159
May 22-27, 2022 ©2022 Association for Computational Linguistics

matching and masked language modeling) to capture text-image alignments. Such general pre-training is inadequate to identify fine-grained aspects, opinions, and their alignments across the language and vision modalities. Therefore, it is important to design task-specific vision-language pre-training, to model aspects, opinions, and their alignments for the MABSA task.

To address this issue, in this paper, we propose a task-specific Vision-Language Pre-training framework for Multimodal Aspect-Based Sentiment Analysis. Specifically, inspired by the recent success of BART-based generative models in text-based ABSA (Yan et al., 2021), we first construct a generative multimodal architecture based on BART (Lewis et al., 2020), for both vision-language pre-training and the downstream MABSA tasks. We then propose three types of vision-language pre-training tasks, including Masked Language Modeling (MLM) and Textual Aspect-Opinion Extraction (AOE) from the language modality, Masked Region Modeling (MRM) and Visual Aspect-Opinion Generation (AOG) from the vision modality, and Multimodal Sentiment Prediction (MSP) across two modalities. Figure 1 illustrates the whole framework of our proposed pre-training approach. Compared with general pre-training methods, our task-specific pre-training approach incorporates multimodal aspect, opinion, and sentiment supervision, which guides pre-trained models to capture important objective and subjective information for the MABSA task.

To evaluate the effectiveness of our pre-training approach, we adopt MVSA-Multi, a widely-used Multimodal Twitter dataset for coarse-grained text-image sentiment analysis (Niu et al., 2016), as our pre-training dataset. We then employ several representative pre-trained models and rule-based methods to obtain the aspect and opinion supervision for our AOE and AOG tasks. As the dataset provides sentiment labels for each multimodal tweet, we adopt them as the supervision for our MSP task.

Our contributions in this work are as follows:

- We introduce a task-specific Vision-Language Pre-training framework for MABSA named VLP-MABSA, which is a unified multimodal encoder-decoder architecture for all the pre-training and downstream tasks.
- Apart from the general MLM and MRM tasks, we further introduce three task-specific pre-training tasks, including Textual Aspect-Opinion Extraction, Visual Aspect-Opinion Generation, and Multimodal Sentiment Prediction, to identify fine-grained aspect, opinions, and their cross-modal alignments.
- Experiments on three MABSA subtasks show that our pre-training approach generally obtains significant performance gains over the state-of-the-art methods. Further analysis on supervised and weakly-supervised settings demonstrates the effectiveness of each pre-training task.

2 Related Work

Vision-Language Pre-training. Inspired by the success of pre-trained language models like BERT (Devlin et al., 2019), many multimodal pre-training models have been proposed (Chen et al., 2020b; Yu et al., 2021; Zhang et al., 2021b) to perform many vision-language tasks which achieve fantastic success. Correspondingly, many general pre-training tasks are proposed, such as Masked Language Modeling (MLM), Masked Region Modeling (MRM) and Image-Text Matching (ITM) (Chen et al., 2020b; Yu et al., 2021). Besides, in order to make the pre-trained models better understand downstream tasks, researchers also design task-specific pre-training models for different downstream tasks (Hao et al., 2020; Xing et al., 2021). In our work, apart from the popular general pre-training tasks, we also design three kinds of task-specific pre-training tasks for the MABSA task.

Text-based Joint Aspect-Sentiment Analysis (JASA). JASA aims to extract aspect terms in the text and predict their sentiment polarities. Many approaches have been proposed including pipeline approaches (Zhang et al., 2015; Hu et al., 2019), multi-task learning approaches (He et al., 2019; Hu et al., 2019) and collapsed label-based approaches (Li et al., 2019; Hu et al., 2019; Chen et al., 2020a). Recently, Yan et al. (2021) proposed a unified generative framework which achieves highly competitive performance on several benchmark datasets for JASA.

Multimodal Sentiment Analysis. Multimodal Sentiment Analysis (MSA) in social media posts is an important direction of sentiment analysis. Many neural network approaches have been proposed to perform the coarse-grained MSA in the literature, which aim to detect the overall sentiment of each input social post (You et al., 2015, 2016; Luo et al., 2017; Xu et al., 2018; Yang et al., 2021b). Different

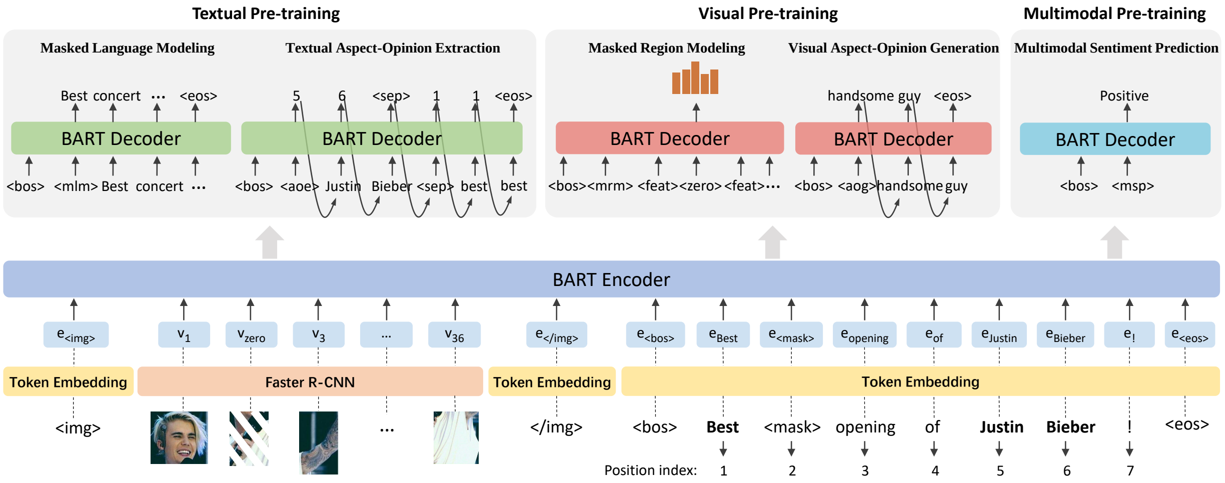

Textual Pre-training
Visual Pre-training
Multimodal Pre-training
Masked Language Modeling
Textual Aspect-Opinion Extraction
Masked Region Modeling
Visual Aspect-Opinion Generation
Multimodal Sentiment Prediction
Best concert ··· <eos>
5 6 <sep> 1 1 <eos>
handsome guy <eos>
Positive
BART Decoder
<bos> <mlm> Best concert ···
<bos> <aoe> Justin Bieber <sep> best best
<bos><mrm> <feat><zero><feat>···
<bos> <aog> handsome guy
<bos> <msp>
BART Encoder
e<img> v1 vzero v3 ··· v36 e</img> e<bos> eBest e<mask> eopening eof eJustin eBieber e! e<eos>
Token Embedding
Faster R-CNN
<img> ··· </img> <bos> Best <mask> opening of Justin Bieber ! <eos>
Position index: 1 2 3 4 5 6 7

Figure 1: Overview of our Vision-Language Pre-Training framework for MABSA

from these studies, our work focuses on the fine-grained MABSA task, which aims to identify the sentiments towards all the aspects mentioned in each input social post.

Multimodal Aspect-Based Sentiment Analysis. As an important sentiment analysis task, many approaches have been approached to tackle the three subtasks of MABSA, including Multimodal Aspect Term Extraction (Zhang et al., 2018; Yu et al., 2020b; Wu et al., 2020a,b; Sun et al., 2020; Zhang et al., 2021a), Multimodal Aspect Sentiment Classification (Xu et al., 2019; Yu et al., 2020a; Yang et al., 2021a; Khan and Fu, 2021) and Joint Multimodal Aspect-Sentiment Analysis (Ju et al., 2021). In this work, we aim to propose a general pre-training framework to improve the performance of all the three subtasks.

3 Methodology

Figure 1 shows the overview of our model architecture. The backbone of our model is BART (Lewis et al., 2020), which is a denoising autoencoder for sequence-to-sequence models. We extend BART to encode both textual and visual inputs, and decode pre-training and downstream tasks from different modalities. In the following subsections, we first introduce our feature extractor, and then illustrate the encoder and decoder of our model, followed by describing the details of three types of pre-training tasks and downstream MABSA tasks.

3.1 Feature Extractor

Image Representation. Following many existing Vision-Language pre-training models (Chen et al., 2020b; Yu et al., 2021), we employ Faster R-CNN (Anderson et al., 2018) to extract visual features. Specifically, we adopt Faster R-CNN to extract all the candidate regions from an input image. We then only retain 36 regions with the highest confidence. Meanwhile, we also keep the semantic class distribution of each region, which will be used for the Masked Region Modeling task. For the retained regions, we use mean-pooled convolutional features processed by Faster R-CNN as our visual features. Let us use $R = \{r_1, ..., r_{36}\}$ to denote the visual features, where $r_i \in \mathbb{R}^{2048}$ refers to the visual feature of the i-th region. To be consistent with the text representation, we adopt a linear transformation layer to project visual features to d-dimensional vectors, denoted by $V \in \mathbb{R}^{d\times 36}$.

Text Representation. For text input, we first tokenize the text and then feed tokens to the embedding matrix. The embeddings of text tokens are used as text features. Let us use $E = \{e_1, ..., e_T\}$ to denote the token indexes of text inputs where T denotes the length of the input text, and $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_T\}$ to denote the embeddings of tokens.

3.2 BART-based Generative Framework

We employ a BART-based generative framework for both vision-language pre-training and downstream MABSA tasks.

Encoder. The encoder of our model is a multi-layer bidirectional Transformer. As shown in Figure 1, to distinguish inputs of different modalities, we follow Xing et al. (2021) by using $\langle img \rangle$ and $\langle /img \rangle$ to indicate the start and the end of visual features, and $\langle bos \rangle$ and $\langle eos \rangle$ to indicate the textual input. In the following part of the paper, we denote the concatenated multimodal input by X.

Decoder. The decoder of our model is also a multi-layer Transformer. The difference is that the

Sentiment	#Image-Text Pairs	#Apects	#Opinions	#Words
Positive	11903	10593	22752	215044
Neutral	4107	3756	7567	74456
Negative	1500	1016	2956	25211

Table 2: The statistics of the MVSA-Multi Dataset. #Apects and #Opinions are the number of aspect terms and opinion terms we extract from the dataset by the rule-based methods introduced in Section 3.3.1.

decoder is unidirectional when generating outputs, while the encoder is bidirectional. Since all pre-training tasks share the same decoder, we insert two special tokens at the beginning of the inputs of the decoder to indicate different pre-training tasks. Following Yan et al. (2021), we insert a special token $\langle bos \rangle$ to indicate the beginning of generation, and then insert a task-specific special token to indicate the task type. Specifically, the special tokens for Masked Language Modeling, Textual Aspect-Opinion Extraction, Masked Region Modeling, Visual Aspect-Opinion Generation, and Multimodal Sentiment Prediction are $\langle bos \rangle\langle mlm \rangle$, $\langle bos \rangle\langle aoe \rangle$, $\langle bos \rangle\langle mrm \rangle$, $\langle bos \rangle\langle aog \rangle$, and $\langle bos \rangle\langle msp \rangle$, respectively.

3.3 Pre-training Tasks

The dataset we use for pre-training is MVSA-Multi (Niu et al., 2016), which is widely used in Multimodal Twitter Sentiment Analysis (Yadav and Vishwakarma, 2020; Yang et al., 2021b). This dataset provides image-text input pairs and coarse-grained sentiments of image-text pairs. Statistics of the dataset are given in Table 2.

With the dataset, we design three types of pre-training tasks, including textual, visual, and multimodal pre-training as follows.

3.3.1 Textual Pre-training

Textual Pre-training contains two tasks: a general Masked Language Modeling task to build alignment between textual and visual features and a task-specific Textual Aspect-Opinion Extraction task to extract aspects and opinions from text.

Masked Language Modeling (MLM). In the MLM pre-training task, we use the same strategy as BERT (Devlin et al., 2019) by randomly masking the input text tokens with a probability of 15%. The goal of the MLM task is to generate the original text based on the image and the masked text, and thus the loss function of the MLM task is:

$$\mathcal{L}_{MLM} = -\mathbb{E}_{X\sim D}\sum_{i=1}^{T}\log P(e_i|e_{<i},\tilde{X}), \quad (1)$$

where e_i and $\tilde{X}$ denote the i^{th} token of the input text and the masked multimodal input, respectively. T is the length of input text.

Textual Aspect-Opinion Extraction (AOE). The AOE task aims to extract aspect and opinion terms from the text. Since the MVSA-Multi dataset does not provide annotations for aspect and opinion terms, we resort to a pre-trained model for aspect extraction and a rule-based method for opinion extraction. Specifically, for aspect extraction, we employ the pre-trained model from a well-known Named Entity Recognition (NER) tool for tweets (Ritter et al., 2011) to perform NER on each tweet in the dataset, and regard the recognized entities as aspect terms. For opinion extraction, we utilize a widely-used sentiment lexicon named SentiWordNet (Esuli and Sebastiani, 2006) to obtain the dictionary of opinion words. Given each tweet, if its sub-sequences (i.e., words or phrases) match the words in the dictionary, we treat them as opinion terms. These extracted aspect and opinion terms are used as the supervision signal of our AOE task.

With the textual aspect-opinion supervision, we follow Yan et al. (2021) by formulating the AOE task as an index generation task. Given the input text as the source sequence, the goal is to generate a target index sequence which consists of the start and end indexes of all aspect and opinion terms. Let us use $Y = [a_1^s, a_1^e, ..., a_M^s, a_M^e, \langle sep \rangle, o_1^s, o_1^e, ..., o_N^s, o_N^e, \langle eos \rangle]$ to denote the target index sequence, where M and N are the number of aspect terms and opinion terms, a^s,a^e and o^s,o^e indicate the start and end indexes of an aspect term and an opinion term respectively, $\langle sep \rangle$ is used to separate aspect terms and opinion terms, and $\langle eos \rangle$ informs the end of extraction. For example, as shown in Figure 1, the extracted aspect and opinion terms are *Justin Bieber* and *best* respectively, and the target sequence is $Y = [5, 6, \langle sep \rangle, 1, 1, \langle eos \rangle]$. For y_t in the target sequence Y, it is either a position index or a special token (e.g., $\langle sep \rangle$). We use $C = [\langle sep \rangle, \langle eos \rangle]$ to denote the set of special tokens, and $\mathbf{C}^d$ as their embeddings.

We assume that $\mathbf{H}^e$ denotes the encoder output of the concatenated multimodal input, $\mathbf{H}_T^e$ denotes the textual part of $\mathbf{H}^e$, and $\mathbf{H}_V^e$ denotes the visual part of $\mathbf{H}^e$. The decoder takes the multimodal encoder output $\mathbf{H}^e$ and the previous decoder output $Y_{<t}$ as inputs, and predicts the token probability

distribution $P(y_t)$ as follows:

$$\mathbf{h}_t^d = \text{Decoder}(\mathbf{H}^e; Y_{<t}), \quad (2)$$

$$\bar{\mathbf{H}}_T^e = (\mathbf{W} + \mathbf{H}_T^e)/2, \quad (3)$$

$$P(y_t) = \text{Softmax}([\bar{\mathbf{H}}_T^e; \mathbf{C}^d]\mathbf{h}_t^d), \quad (4)$$

where $\mathbf{W}$ denotes the embeddings of input tokens. The loss function of the AOE task is as follows:

$$\mathcal{L}_{AOE} = -\mathbb{E}_{X \sim D} \sum_{t=1}^{O} \log P(y_t | Y_{<t}, X), \quad (5)$$

where $O = 2M + 2N + 2$ is the length of Y and X denotes the multimodal input.

3.3.2 Visual Pre-training

Visual Pre-training contains two tasks: a general Masked Region Modeling task and a task-specific Visual Aspect-Opinion Generation task to capture subjective and objective information in the image.

Masked Region Modeling (MRM). Following Xing et al. (2021), our MRM task aims to predict the semantic class distribution of the masked region. As shown in Figure 1, for the input of the encoder, we randomly mask image regions with a probability of 15%, which are replaced with zero vectors. For the input of the decoder, we first add two special tokens ⟨*bos*⟩⟨*mrm*⟩, and then represent each masked region with ⟨*zero*⟩ and each remaining region with ⟨*feat*⟩. After feeding the input to the decoder, an MLP classifier is stacked over the output of each ⟨*zero*⟩ to predict the semantic class distribution. Let us use $p(v_z)$ to denote the predicted class distribution of the z-th masked region, and $q(v_z)$ to denote the class distribution detected by Faster R-CNN. The loss function for MRM is to minimize the KL divergence of the two class distributions:

$$\mathcal{L}_{MRM} = \mathbb{E}_{X \sim D} \sum_{z=1}^{Z} D_{KL}(q(v_z) || p(v_z)), \quad (6)$$

where Z is the number of masked regions.

Visual Aspect-Opinion Generation (AOG). The AOG task aims to generate the aspect-opinion pair detected from the input image. In the field of Computer Vision, Borth et al. (2013) proposed to detect the visual sentiment concept, i.e., Adjective-Noun Pair (ANP) such as *smiling man* and *beautiful landscape* in the image. Since the nouns and adjectives of ANP respectively capture the fine-grained aspects and opinions in the image, we regard ANPs as visual aspect-opinion pairs. In order to detect the ANP of each input image, we adopt a pre-trained ANP detector DeepSentiBank[2] (Chen et al., 2014) to predict the class distribution over 2089 pre-defined ANPs. The ANP with the highest probability is selected as the supervision signal of our AOG task. For example, in Figure 1, the ANP detected from the input image is *handsome guy*, and we regard it as the supervision.

With the visual aspect-opinion supervision, we formulate the AOG task as a sequence generation task. Specifically, let us use $G = \{g_1, ..., g_{|G|}\}$ to denote the tokens of the target ANP and $|G|$ to denote the number of ANP tokens. The decoder then takes the multimodal encoder output $\mathbf{H}^e$ and the previous decoder output $G_{<i}$ as inputs, and predicts the token probability distribution $P(g_i)$:

$$\mathbf{h}_i^d = \text{Decoder}(\mathbf{H}^e; G_{<i}), \quad (7)$$

$$P(g_i) = \text{Softmax}(\mathbf{E}^T \mathbf{h}_i^d), \quad (8)$$

where $\mathbf{E}$ denotes the embedding matrix of all tokens in the vocabulary.

The loss function of the AOG task is:

$$\mathcal{L}_{AOG} = -\mathbb{E}_{X \sim D} \sum_{i=1}^{|G|} \log P(g_i | g_{<i}, X). \quad (9)$$

3.3.3 Multimodal Pre-training

Multimodal Pre-training has one task named Multimodal Sentiment Prediction (MSP). Different from the aforementioned pre-training tasks whose supervision signals only come from one modality, the supervision signals for MSP come from multimodality, which can enhance models to identify the subjective information in both language and vision and capture their rich alignments.

Multimodal Sentiment Prediction (MSP). As the MVSA-Multi dataset provides the coarse-grained sentiment labels for all the text-image pairs, we use the sentiment labels as supervision signals of our MSP task. Formally, we model the MSP task as a classification task, where we first feed the two special tokens ⟨*bos*⟩⟨*msp*⟩ to the decoder and then predict the sentiment distribution $P(s)$ as follows:

$$\mathbf{h}_{msp}^d = \text{Decoder}(\mathbf{H}^e; \mathbf{E}_{msp}), \quad (10)$$

$$P(s) = \text{Softmax}(\text{MLP}(\mathbf{h}_{msp}^d)), \quad (11)$$

where $\mathbf{E}_{msp}$ is the embeddings of two special tokens.

[2] https://github.com/stephen-pilli/DeepSentiBank

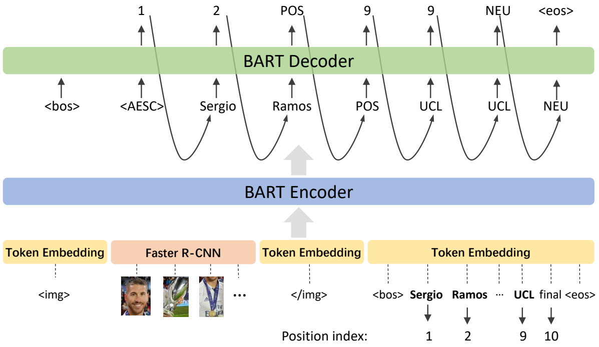


Figure 2: An example of downstream task JMASA. $\langle AESC \rangle$ informs the current task is JMASA.

We use the cross-entropy loss for the MSP task:

$$\mathcal{L}_{MSP} = -\mathbb{E}_{X \sim D} \log P(s|X), \qquad (12)$$

where s is the golden sentiment annotated in dataset.

3.3.4 Full Pre-training Loss

To optimize all the model parameters, we adopt the alternating optimization strategy to iteratively optimize our five pre-training tasks. The objective function is as follows:

$$\begin{aligned}\mathcal{L} =& \lambda_1 \mathcal{L}_{MLM} + \lambda_2 \mathcal{L}_{AOE} + \lambda_3 \mathcal{L}_{MRM} + \\ & \lambda_4 \mathcal{L}_{AOG} + \lambda_5 \mathcal{L}_{MSP}\end{aligned} \qquad (13)$$

where λ_1, λ_2, λ_3, λ_4, and λ_5 are tradeoff hyper-parameters to control the contribution of each task.

3.4 Downstream Tasks

We consider all the three subtasks in MABSA as our downstream tasks, including Joint Multimodal Aspect-Sentiment Analysis (JMASA), Multimodal Aspect Term Extraction (MATE), and Multimodal Aspect-oriented Sentiment Classification (MASC). We model these downstream tasks based on the same BART-based generative framework in vision-language pre-training, so that the downstream task can benefit more from pre-training during the fine-tuning stage. Following Yan et al. (2021), we formulate the outputs of the three subtasks as follows:

- JMASA: $Y = [a_1^s, a_1^e, s_1, ..., a_i^s, a_i^e, s_i, ...]$,
- MATE: $Y = [a_1^s, a_1^e, ..., a_i^s, a_i^e, ...]$,
- MASC: $Y = [\underline{a_1^s}, \underline{a_1^e}, s_1, ..., \underline{a_i^s}, \underline{a_i^e}, s_i, ...]$,

where a_i^s, a_i^e, and s_i inform the start index, end index, and sentiment of an aspect term in the text. The underlined tokens are given during inference.

	TWITTER-2015			TWITTER-2017		
	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 Aspects	3179	1122	1037	3562	1176	1234
#Sentence	2101	727	674	1746	577	587

Table 3: The basic statistics of two TWITTER datasets.

Similar to the AOE task in Section 3.3.1, we formulate all the subtasks as index generation tasks, and use Eqn. (2) to Eqn. (4) to generate the token distribution. The difference is that the special token set is modified as $C = [\langle POS \rangle, \langle NEU \rangle, \langle NEG \rangle, \langle EOS \rangle]$ by adding the sentiment categories. Figure 2 shows an example for JMASA. Since the aspect-sentiment pairs are (*Sergio Ramos*, *Positive*) and (*UCL*, *Neutral*), its target sequence is $[1, 2, \langle POS \rangle, 9, 9, \langle NEU \rangle, \langle eos \rangle]$.

4 Experiment

4.1 Settings

Downstream datsets. We adopt two benchmark datasets annotated by Yu and Jiang (2019), namely TWITTER-2015 and TWITTER-2017 to evaluate our model. The statistics of the two datasets are shown in Table 3.

Implementation Details. We employ BART-base (Lewis et al., 2020) as our framework. Specifically, the encoder and decoder both have six layers and are initialized with BART-base parameters. We fix all the hyper-parameters after tuning them on the development set. The pre-training tasks were trained for 40 epochs and the downstream tasks were fine-tuned for 35 epochs. The batch sizes are set to 64 and 16, respectively. The learning rate is set to 5e-5. The hidden size of our model is set to 768, which is the same as BART. The tradeoff hyper-parameters λ_1, λ_2, λ_3, λ_4, and λ_5 are all set to 1. Note that for the subtask MASC, different from Ju et al. (2021) evaluating on the correctly predicted aspects, we provide all the golden aspects to the decoder of our framework during the inference stage and evaluate on all the aspects. We implement all the models with PyTorch, and run experiments on a RTX3090 GPU.

Evaluation Metrics. We evaluate our model over three subtasks of MABSA and adopt Micro-F1 score (F1), Precision (P) and Recall (R) as the evaluation metrics to measure the performance. For MASC, to fairly compare with other approaches,

	TWITTER-2015			TWITTER-2017		
	P	**R**	**F1**	**P**	**R**	**F1**
Text-based methods						
SPAN*	53.7	53.9	53.8	59.6	61.7	60.6
D-GCN*	58.3	58.8	59.4	64.2	64.1	64.1
BART	62.9	65.0	63.9	65.2	65.6	65.4
Multimodal methods						
UMT+TomBERT*	58.4	61.3	59.8	62.3	62.4	62.4
OSCGA+TomBERT*	61.7	63.4	62.5	63.4	64.0	63.7
OSCGA-collapse*	63.1	63.7	63.2	63.5	63.5	63.5
RpBERT-collapse*	49.3	46.9	48.0	57.0	55.4	56.2
JML*	65.0	63.2	64.1	66.5	65.5	66.0
VLP-MABSA	**65.1**	**68.3**	**66.6**	**66.9**	**69.2**	**68.0**

Table 4: Results of different approaches for JMASA. * denotes the results are from Ju et al. (2021).

we also use Accuracy (Acc).

4.2 Compared Systems

In this section, we introduce four types of compared systems for different tasks.

Approaches for Multimodal Aspect Term Extraction (MATE). 1) *RAN* (Wu et al., 2020a), which aligns text with object regions by a co-attention network. 2) *UMT* (Yu et al., 2020b), which uses Cross-Modal Transformer to fuse text and image representations for Multimodal Named Entity Recognition (MNER). 3) *OSCGA* (Wu et al., 2020b), another MNER approach using visual objects as image representations. 4) *RpBERT* (Sun et al., 2021), which uses a multitask training model for MNER and image-text relation detection.

Approaches for Multimodal Aspect Sentiment Classification (MASC). 1) *TomBERT* (Yu and Jiang, 2019), which tackles the MASC task by employing BERT to capture intra-modality dynamics. 2) *CapTrBERT* (Khan and Fu, 2021), which translates the image to a caption as an auxiliary sentence for sentiment classification.

Text-based approaches for Joint Aspect-Sentiment Analysis (JASA). 1) *SPAN* (Hu et al., 2019), which formulates the JASA task as a span prediction problem. 2) *D-GCN* (Chen et al., 2020a), which proposes a directional graph convolutional network to capture the correlation between words. 3) *BART* (Yan et al., 2021), which adapts the JASA task to BART by formulating it as an index generation problem.

Multimodal approaches for Joint Multimodal Aspect-Sentiment Analysis (JMASA). 1) *UMT+TomBERT* and *OSCGA+TomBERT*, which are simple pipeline approaches by combining methods for subtasks mentioned above. 2) *UMT-collapsed* (Yu et al., 2020b), *OSCGA-collapsed* (Wu et al., 2020b) and *RpBERT-collapsed* (Sun et al., 2021), which model the JMASA task with collapsed labels such as *B-POS* and *I-POS*. 3) *JML* (Ju et al., 2021), which is a multi-task learning approach proposed recently with the auxiliary cross-modal relation detection task.

Methods	TWITTER-2015			TWITTER-2017		
	P	**R**	**F1**	**P**	**R**	**F1**
RAN*	80.5	81.5	81.0	90.7	90.7	90.0
UMT*	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA*	81.7	82.1	81.9	90.2	90.7	90.4
JML-MATE*	83.6	81.2	82.4	**92.0**	90.7	91.4
VLP-MABSA	**83.6**	**87.9**	**85.7**	90.8	**92.6**	**91.7**

Table 5: Results of different approaches for MATE. * denotes the results are from Ju et al. (2021).

Methods	TWITTER-2015		TWITTER-2017	
	Acc	**F1**	**Acc**	**F1**
TomBERT	77.2	71.8	70.5	68.0
CapTrBERT	78.0	73.2	72.3	70.2
JML-MASC	**78.7**	-	72.7	-
VLP-MABSA	78.6	**73.8**	**73.8**	**71.8**

Table 6: Results of different approaches for MASC. Note that JML-MASC only evaluates on the aspects correctly predicted by JML-MATE while the other methods evaluate on all the golden aspects.

4.3 Main Results

In this section, we analyze the results of different approaches on three subtasks of MABSA.

Results of JMASA. Table 4 shows the results of different methods for JMASA. As we can see from the table, *BART* achieves the best performance among text-based methods, and it even outperforms some multimodal methods, which proves the superiority of our base framework. For multimodal methods, *JML* achieves better performance than previous methods mainly due to its auxiliary task about relation detection between image and text. Among all the methods, *VLP-MABSA* which is the whole model with all the pre-training tasks consistently performs the best across two datasets. Specifically, it significantly outperforms the second best system *JML* with 2.5 and 2.0 absolute percentage points with respect to *F1* on TWITTER-2015 and TWITTER-2017, respectively. This mainly benefits from our task-specific pre-training tasks, which identify aspects and opinions as well as their alignments across the two modalities.

Results of MATE and MASC. Table 5 and Table 6 show the results of MATE and MASC, re-

		TWITTER-2015			TWITTER-2017		
		JMASA	MATE	MASC	JMASA	MATE	MASC
Full supervision	w/o pre-training	65.31	84.80	76.81	66.10	90.67	72.78
	+$\mathbf{T}_{\text{MLM}}$	65.44	84.91	77.08	66.27	91.00	72.82
	+$\mathbf{T}_{\text{AOE}}$	65.92	85.43	77.48	67.12	91.75	72.89
	+$\mathbf{V}_{\text{MRM}}$	65.94	85.49	77.53	67.15	91.72	73.13
	+$\mathbf{V}_{\text{AOG}}$	66.38	**85.73**	77.82	67.66	**91.77**	73.32
	+ $\mathbf{MM}_{\text{MSP}}$	**66.64**	85.66	**78.59**	**68.05**	91.73	**73.82**
Weak supervision	w/o pre-training	39.79	69.33	57.40	49.12	80.48	61.04
	+$\mathbf{T}_{\text{MLM}}$	40.42	69.69	58.00	49.69	81.26	61.15
	+$\mathbf{T}_{\text{AOE}}$	46.15	79.13	58.32	52.00	84.60	61.46
	+$\mathbf{V}_{\text{MRM}}$	46.64	79.49	58.68	52.18	84.47	61.78
	+$\mathbf{V}_{\text{AOG}}$	47.79	**80.94**	59.32	53.16	**85.04**	62.51
	+ $\mathbf{MM}_{\text{MSP}}$	**51.71**	80.69	**62.58**	**55.38**	84.88	**64.42**

Table 7: The results of pre-training tasks on two benchmarks. We evaluate over three tasks JMASA, MATE, and MASC in terms of *F1*, *F1* and *Acc*, respectively. *T*, *V*, and *MM* denote the Textual, Visual, and Multimodal pre-training, respectively. Each row adds an extra pre-training task to the row above it.

spectively. Similar to the trend on the JMASA subtask, we can clearly observe that our proposed approach *VLP-MABSA* generally achieves the best performance across the two datasets, except on the accuracy metric of TWITTER-2015. These observations further demonstrate the general effectiveness of our proposed pre-training approach.

4.4 In-depth Analysis of Pre-training Tasks

To explore the impact of each pre-training task, we perform a thorough ablation study over the full supervision setting which uses full training dataset and the weak supervision setting which only randomly chooses 200 training samples for fine-tuning.

Impact of Each Pre-training Task. As we can see from Table 7, the performance generally improves with respect to most metrics when adding more pre-training tasks.

To better analyze the effect of each pre-training task, we take the weak supervision experiments on TWITTER-2015 as an example. When only using MLM to pre-train our model, the performance only gets slight improvements. After adding the AOE task, the result of MATE gets a huge improvement of 9.44% on *F1*. This shows that the AOE task greatly enhances our model's ability to recognize the aspect terms. When adding the MRM task, the performance gets slight improvements again. This reflects that general pre-training tasks (e.g., MLM and MRM) are not adequate for our model to tackle downstream tasks which need the model to understand the subjective and objective information from image and text. When adding the AOG task, the performance over three subtasks gets a moderate improvement, which proves the effectiveness of the AOG task. Finally, adding the MSP task significantly boosts the performance, especially on the MASC task. This shows that the MSP task can enhance our model's understanding of sentiment across language and image modalities. By combining all the pre-training tasks, our full model generally achieves the best results over most of the subtasks whether in both full supervision and weak supervision settings.

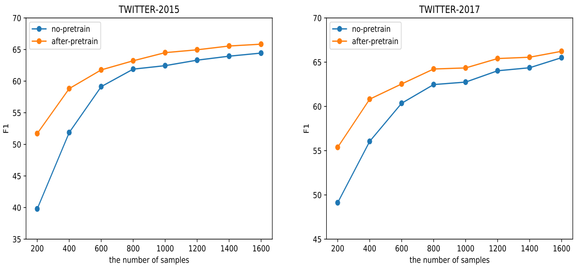


Figure 3: The effectiveness of pre-training when using different number of training samples for the downstream task. Y-axis refers to the **F1** score (%) of the JMASA task.

Impact of pre-training when using different number of downstream training samples. To better understand the impact of pre-training, we compare the results with and without pre-training when adopting different number of samples for downstream training. We use the JMASA task as the example to observe the impact. As shown in Fig. 3, when the sample size is small, pre-training can bring a huge improvement. In contrast, when the sample size becomes larger, pre-training brings relatively small improvements. This further illustrates the robustness and the effectiveness of our pre-training approach, especially in low-resource scenarios.

4.5 Case Study

To further demonstrate the effectiveness of our approach, we present four test examples with predictions from different methods. The compared methods are *BART*, our framework using multimodal inputs without pre-training (denoted by *MM*), and our framework using multimodal inputs with full pre-training (denoted by *VLP*), respectively. As shown in Table 8, for example (a), both *BART* and *MM* extracted the wrong aspect term (i.e., *the Faithfull Pearl Jam*) and gave the incorrect sentiment prediction towards *Eddie*. For example (b), *BART* only extracted one aspect term *Madonna* while *MM* identified an additional aspect term *Demelza*. However, the sentiment towards *Madonna* was wrongly predicted by *MM*. For example (c), *BART* only

Image				
Text	(a) RT @ PearlJam : Eddie and the Faithfull Pearl Jam fans in Buenos Aires . Photo by @ epozzoni # PJSA2013	(b) RT @ BBCOne : Dear Madonna , THIS is how you wear a cape . # Poldark # Demelza	(c) RT @ TrumpDoral : Congratulations to the the new # MissUniverse , Miss Colombia , Paulina Vega !	(d) RT @ myfox8 : Charlotte @ hornets visit # Greensboro for D - League meeting
GT	(Eddie, POS) (Pearl Jam, POS) (Buenos Aires, NEU)	(Madonna, POS) (Poldark, NEU) (Demelza, NEU)	(Miss Colombia, POS) (Paulina Vega, POS)	(Charlotte, NEU) (Greensboro, NEU) (D – League, NEU)
BART	(Eddie, NEU) × (the Faithfull Pearl Jam, NEU) × (Buenos Aires, NEU) ✓	(Madonna, POS) ✓ - × - ×	(Colombia, POS) × (Paulina Vega, POS) ✓	(Charlotte, NEU) ✓ (Greensboro, NEU) ✓ - ×
MM	(Eddie, NEU) × (the Faithfull Pearl Jam, NEU) × (Buenos Aires, NEU) ✓	(Madonna, NEU) × - × (Demelza, NEU) ✓	(Colombia, NEU) × (Paulina Vega, POS) ✓	(Charlotte, NEU) ✓ (Greensboro, NEU) ✓ - ×
VLP	(Eddie, POS) ✓ (Pearl Jam, POS) ✓ (Buenos Aires, NEU) ✓	(Madonna, POS) ✓ (Poldark, NEU) ✓ (Demelza, NEU) ✓	(Miss Colombia, POS) ✓ (Paulina Vega, POS) ✓	(Charlotte, NEU) ✓ (Greensboro, NEU) ✓ (D – League, NEU) ✓

Table 8: Predictions of different methods on four test samples. NEU, POS, and NEG denote Neutral, Positive, and Negative sentiments, respectively.

recognized part of the aspect term *Colombia* and *MM* wrongly predicted the sentiment towards *Miss Colombia* as *Neutral*. For example (d), both *BART* and *MM* failed to recognize the aspect term *D-League*. Among all the cases, our *VLP* model with full pre-training correctly extracted all the aspect terms and classified the sentiment , which shows the advantage of our generative framework and task-specific pre-training tasks.

5 Conclusion

In this paper, we proposed a task-specific Vision-Language Pre-training framework for Multimodal Aspect-Based Sentiment Analysis (VLP-MABSA). We further designed three kinds of pre-training tasks from the language, vision, and multi-modal modalities, respectively. Experimental results show that our proposed approach generally outperforms the state-of-the-art methods for three subtasks of MABSA. Our work is a first step towards a unified Vision-Language Pre-training framework for MABSA. In the future, we plan to apply our pre-training approach on a larger dataset and consider the relation between image and text in our pre-training framework. We hope this work can potentially bring new insights and perspectives to the research of MABSA.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments. This work was supported by the Natural Science Foundation of China (62076133 and 62006117), and the Natural Science Foundation of Jiangsu Province for Young Scholars (BK20200463) and Distinguished Young Scholars (BK20200018).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232.

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 272–279. International Committee on Computational Linguistics.

Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of ACL*, pages 504–515.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of ACL*, pages 537–546.

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 EMNLP*.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentimentclassification through input space translation. In *Proceedings of ACM MM*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of ACL*, pages 7871–7880.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6714–6721.

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 ACL*, pages 1990–1999.

Jiebo Luo, Damian Borth, and Quanzeng You. 2017. Social multimedia sentiment analysis. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1953–1954.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*, pages 15–27. Springer.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *EMNLP*.

Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. Riva: A pre-trained tweet multimodal model based on text-image relation for multimodal ner. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1852–1862.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13860–13868.

Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020a. Multimodal aspect extraction with region-aware alignment network. In *Proceedings of NLPCC*.

Zhiwei Wu, Changmeng Zheng, Cai Yi, Leung Hofung Chen Junying, and Qing Li. 2020b. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of ACM MM*.

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. Kmbart: Knowledge enhanced multimodal bart for visual commonsense generation. In *Proceedings of ACL*.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A comemory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 929–932.

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

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385.

Hang Yan, Junqi Dai, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of ACL-IJCNLP*, pages 2416–2429.

Li Yang, Jianfei Yu, Chengzhi Zhang, and Jin-Cheon Na. 2021a. Fine-grained sentiment analysis of political tweets with entity-aware multimodal network. In *International Conference on Information*, pages 411–420.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021b. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of ACL-IJCNLP*, pages 328–339.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Joint visual-textual sentiment analysis with deep neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1071–1074.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM international conference on Web search and data mining*, pages 13–22.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3208–3216.

Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of IJCAI*, pages 5408–5414.

Jianfei Yu, Jing Jiang, and Rui Xia. 2020a. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020b. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of ACL*.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14347–14355.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.