Dual-Encoder Transformers with Cross-modal Alignment for Multimodal Aspect-based Sentiment Analysis

Zhewen Yu[†], **Jin Wang**[†*], **Liang-Chih Yu**[‡*] and **Xuejie Zhang**[†]
[†]School of Information Science and Engineering, Yunnan University, Yunnan, P.R. China
[‡]Department of Information Management, Yuan Ze University, Taiwan
Contact:wangjin@ynu.edu.cn, lcyu@saturn.yzu.edu.tw

Abstract

Multimodal aspect-based sentiment analysis (MABSA) aims to extract the aspect terms from text and image pairs, and then analyze their corresponding sentiment. Recent studies typically use either a pipeline method or a unified transformer based on a cross-attention mechanism. However, these methods fail to explicitly and effectively incorporate the alignment between text and image. Supervised finetuning of the universal transformers for MABSA still requires a certain number of aligned image-text pairs. This study proposes a dual-encoder transformer with cross-modal alignment (DTCA). Two auxiliary tasks, including text-only extraction and text-patch alignment are introduced to enhance cross-attention performance. To align text and image, we propose an unsupervised approach which minimizes the Wasserstein distance between both modalities, forcing both encoders to produce more appropriate representations for the final extraction. Experimental results on two benchmarks demonstrate that DTCA consistently outperforms existing methods. For reproducibility, the code for this paper is available at: https://github.com/windforfurture/DTCA.

1 Introduction

Human experience of the world is multimodal, e.g., seeing objects, hearing sounds, feeling textures, and tasting flavors. Multimodal experiences are usually mutually associated to some extent. For example, images are usually associated with tags and text explanations, and text often contains images to more clearly express the main intent of the author.

With the widespread availability of smart phones with digital cameras, social media posts have become increasingly multimodal . To practically apply the existing aspect-based sentiment analysis, one must be able to interpret such multimodal attributes together (Yu et al., 2022; Ling et al., 2022).

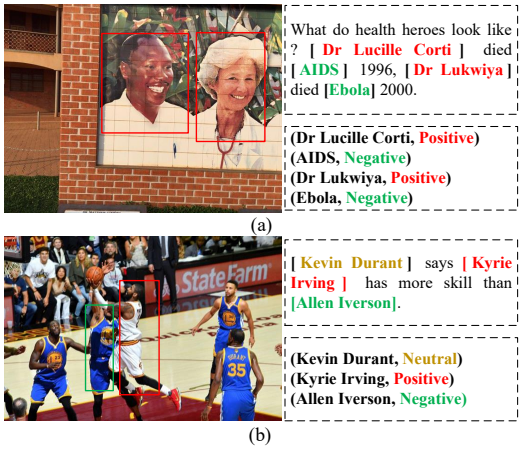


Figure 1: Two examples of joint multimodal aspect sentiment analysis.

Figure 1 (a) shows an example: *What do health heroes look like? Dr Lucille Corti died AIDS 1996, Dr Lukwiya died Ebola 2000.* An intelligent system is expected to extract four aspect-sentiment pairs from this text, i.e., (*Dr Lucille Corti*, **positive**), (*AIDS*, **negative**), (*Dr Lukwiya*, **positive**) and (*Ebola*, **negative**). Notably, if only the language modality is used for inference, the model tends to predict (*Dr Lucille Corti*, **negative**) and (*Dr Lukwiya*, **negative**). Related to the vision modality, the expression of the text will become more ironic, and thus tends to be positive. Figure 1 (b) shows another example: *Kevin Durant says Kyrie Irving has more skill than Allen Iverson*. It is difficult to infer from the image that this person is necessarily good at basketball, while a direct understanding of the text seems to recognize the attitude of the author towards *Kyrie Irving and Allen Iverson*.

Based on this, existing methods for multimodal aspect-based sentiment analysis are typically composed of two subtasks in a pipeline model, including multimodal aspect term extraction (MATE) and multimodal aspect sentiment classification (MASC). The former tries to identify all the as-

*Corresponding authors.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 414–423
November 20–23, 2022.

pect terms from texts (Wang et al., 2021), while the latter aims to classify the sentiment for each identified aspect term (Hosseini-Asl et al., 2022; Zhang et al., 2021b; Yuan et al., 2022). Unfortunately, the pipeline approach ignores the innate relationship between the two subtasks and is prone to error propagation.

Alternatively, another obvious solution is to apply multitasked learning to integrate both subtasks into a joint framework (Vazan and Razmara, 2021). Combining different modalities or types of information to improve performance seems intuitively appealing, but it is challenging in practice to reconcile the varying levels of noise and conflicts between modalities. A series of convolution-based models are usually applied to extract image features, including VGG (Simonyan and Zisserman, 2015) and ResNet (He et al., 2015). To extract region-of-interest (ROI) features, several subsequent works have used a Fast R-CNN (Girshick, 2015) to learn the image representation (Zhang et al., 2021a). For text, Transformer-based models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) , XLNet (Yang et al., 2019) and ELECTRA (Clark et al., 2020) have greatly improved the capability of language understanding and generation.

Taking the obtained representation of both modalities as input, recent studies applied different attentions to compose the features for the final classification. For examples, Ju et al. (2021) and Xu et al. (2022) applied a cross-modal self-attention approach to learn text-image interaction and obtain image-aware text representations and text-aware image representations. However, the image-text pairs present different kinds of knowledge. Thus, different modalities may contribute differently to the final classification, and do not have equivalent amounts of information in each modality, with the language modality tending to dominate with more information. For training, the gradients from the dominant modality will overwhelm the other, effectively preventing the entire model from being trained. It is difficult to encode explicit cross-modal information by superficially measuring the attention distribution.

Based on the universal Transformer architecture, the unified vision-and-language pretrained models can simultaneously encode both modalities, e.g., OSCAR (Li et al., 2020) and UNITER (Chen et al., 2020). However, they are insensitive to aspect extraction and sentiment detection from both language and vision modalities. Finetuning these models with a supervised learning still require a certain number of aligned image-text pairs.

In this study, a dual-encoder transformer with cross-modal alignment (DTCA) is proposed for multimodal aspect-based sentiment analysis. Instead of extracting ROI features, we apply the ViT strategy (Dosovitskiy et al., 2021), which tokenizes the image by slicing it into a sequence of patches. Both ViT and RoBERTa are initialized from pretrained checkpoints, and were used to encode the vision and language modalities. To align the learned features, a multitask learning architecture containing three subtasks was applied, including text-only extraction, co-attention interaction, and token-patch matching. Aside from the co-attention module, we propose minimizing the Wasserstein distance between tokens and images to improve the training effectiveness of the proposed model.

Comparative experiments were conducted on two different benchmarks. The empirical results show that the proposed model outperforms the existing unimodal and multimodal models for MABSA tasks. The effects on different subtasks were further evaluated, finding that the different subtasks all play an indispensable role in performance improvement.

The remainder of this paper is organized as follows. Section 2 presents a detailed description of the proposed DTCA model. Section 3 summarizes the implementation details and experimental results. Conclusions are drawn in Section 4.

2 Dual-Encoder Transformers

Figure 2 shows the overall architecture of the proposed dual-encoder transformers with cross-modal alignment. Two individual transformer-based models, i.e., RoBERTa (Liu et al., 2019) and ViT (Dosovitskiy et al., 2021), were respectively applied for text and image encoding. Notably, both RoBERTa and ViT share the same encoder architecture, which is initialized from a well pretrained checkpoint. Three subtasks were applied for cross-modal alignment to enhance the performance of cross-modal attention for MABSA.

2.1 Modality-specific Encoder

Tokenizer. An input sample $\mathbf{x}$ consists of two modalities, including an image $\mathbf{v}$ and a text $\mathbf{s}$.The objective of MABSA is to perform sequence la-

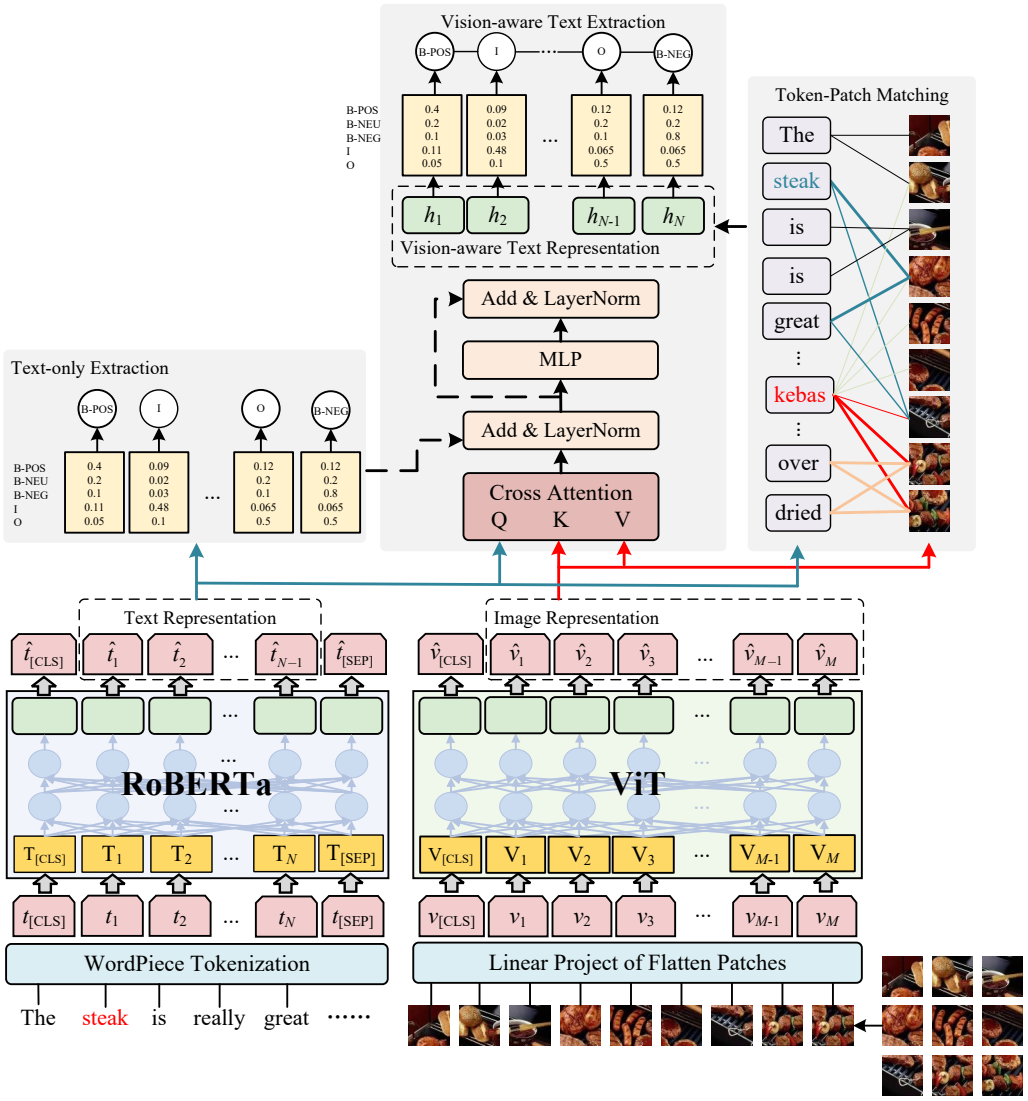


Figure 2: The overall architecture of the proposed dual-encoder Transformers with cross-modal alignment for MABSA.

beling to predict the labels $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ where N is the length of the text. Following the ViT, the image was first sliced into a sequence of patches $\mathbf{v} = [v_1, v_2, \ldots, v_M] \in \Re^{M \times (P^2 \times C)}$, where (P, P) is the resolution of each patch, C is the number of channels, and $M = HW/P^2$ is the resulting number of patches. Each patch was then flattened and prepended with a special token, i.e., $v_{[\text{CLS}]}$, followed by a linear projection $V \in \Re^{(P^2 \times C) \times d_h}$. The result patch embeddings $\bar{v} \in \Re^{(M+1) \times d_h}$ can be formulated as,

$$\bar{v} = [v_{[\text{CLS}]}, v_1 V, v_2 V, ..., v_M V] + V^{pos} \quad (1)$$

where d_h is the dimensionality and $V^{pos} \in \Re^{(M+1) \times d_h}$ is the position embeddings.

For language modality, the input text is tokenized by the WordPiece (Wu et al., 2016) tokenizer as same as in the RoBERTa model to obtain a sequence of token embeddings $\bar{t} \in \Re^{(N+1) \times d_h}$ with a word embedding matrix $T \in \Re^{N \times |\hat{V}|}$ as follows,

$$\bar{t} = [t_{[\text{CLS}]}, t_1 T, t_2 T, ..., t_N T, t_{[\text{SEP}]}] + T^{pos} + T^{seg} \quad (2)$$

where $T^{pos} \in \Re^{(N+1) \times d_h}$ and $T^{seg} \in \Re^{(N+1) \times d_h}$ are respectively the position and segment embeddings, and $|\hat{V}|$ is the number of the vocabulary items. Here, the [CLS] and [SEP] tokens respectively respond to <*s*> and </*s*> tokens in the RoBERTa model. We did not apply any extra embeddings to annotate the type of modality, since

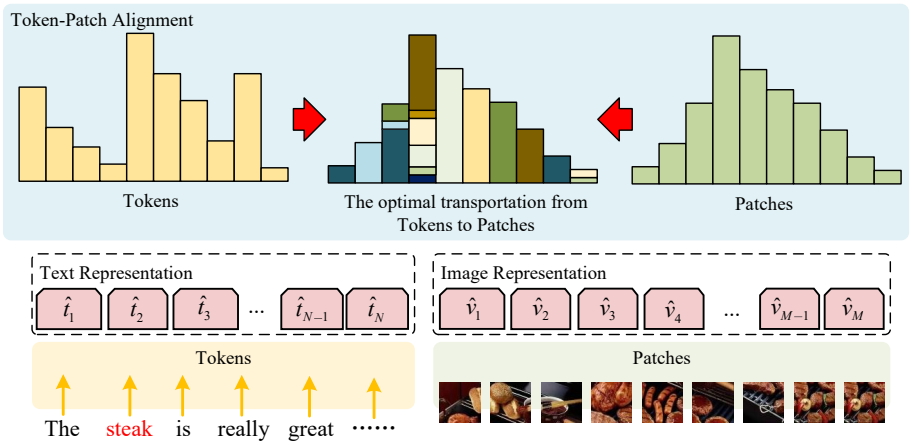


Figure 3: The conceptual diagram of the proposed Token-Patch Alignment.

doing so brings no additional improvement to the proposed model.

Encoders. Both RoBERTa and ViT consist of stacked Transformer blocks including a multi-head self-attention (MHSA) layer and an MLP layer. The MLP consists of two dense connection layers with a GELU non-linear activation. Before both MHSA and MLP, layer normalization (LayerNorm) was applied, which can be formulated as,

$$z^{(0)} = \bar{v} \text{ or } \bar{t} \tag{3}$$

$$\tilde{z}^{(l)} = \text{MHSA}(\text{LayerNorm}(z^{(l-1)})) + z^{(l-1)} \tag{4}$$

$$z^{(l)} = \text{MLP}(\text{LayerNorm}(\tilde{z}^{(l)})) + \tilde{z}^{(l)} \tag{5}$$

where l is the index of the layer of RoBERTa or ViT. The final output of transformer encoder is a hidden representation $z_V^{(L)} = [\hat{v}_1, \hat{v}_2, ..., \hat{v}_M]$ and $z_T^{(L)} = [\hat{t}_1, \hat{t}_2, ..., \hat{t}_N]$ at the last, i.e., the L-th layer, which is used for multitask learning and the final extraction.

For all experiments, the weights of RoBERTa and ViT were respectively initialized from pretrained `roberta-base` and `vit-base-patch16-224-in21k`. The hidden size d_h is 768, the number of layers of encoder L is 12, patch size P is 14, MLP size is 3,072 and the number of attention heads is 12.

2.2 Cross-modal Alignment

To align the features of both the vision and language modalities, we propose a cross-modal alignment to train both the image and text encoders for the final cross-modal extraction. It mainly consists of three subtasks: text-only extraction, co-attention interaction, and token-patch matching.

Text-only Extraction. The textual representation obtained from RoBERTa, i.e., $z_T^{(L)} = [\hat{t}_{[\text{CLS}]}, \hat{t}_1, \hat{t}_2, ..., \hat{t}_N, \hat{t}_{[\text{SEP}]}]$ was fed to a fully-connected layer with softmax activation to predict the auxiliary tags for the tokens, defined as,

$$\hat{\mathbf{y}}_n = \text{softmax}(W^t \hat{t}_n + b^t) \tag{6}$$

where $W^t \in \Re^{K \times d_h}$ and $b^t \in \Re^K$ are trainable parameters, and K is the number of the candidate tags. Given a training dataset of $\{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}_{j=1}^J$, the loss function is a categorical cross-entropy,

$$\mathcal{L}^{TO} = -\frac{1}{J \times N} \sum_{j=1}^{J} \sum_{n=1}^{N} \mathbb{I}(y_n^{(j)}) \circ \log \hat{\mathbf{y}}_n^{(j)} \tag{7}$$

where $y_n^{(j)}$ is the ground-truth label, $\mathbb{I}(y_n)$ denotes a one-hot vector with the y-th component being one, and $\circ$ represents the element-wised multiplication operation.

For token classification, BIO schema was applied. Instead of using 7 tags as in previous works, we used only 5 tags, i.e., `B-POS`, `B-NEU`, `B-NEG`, `I` and `O`. For example, the sequence of `{B-POS, I-POS}` can be converted to `{B-POS, I}`, so that the number of class K can be compressed by half, thus decrease the prediction error caused by sentiment analysis.

Vision-aware Text Extraction. Multi-head cross-attention was applied to integrate the textual and visual features, where the text representation $z_T^{(L)} = [\hat{t}_1, \hat{t}_2, ..., \hat{t}_N]$ is regarded as the query, while the image representation $z_V^{(L)} = [\hat{v}_1, \hat{v}_2, ..., \hat{v}_M]$ was

Datasets		#S	#A	#Pos	#Neu	#Neg	MA	MS	Mean	Max
Twitter-2015	Train	2100	3179	928	1883	368	800	278	15	35
	Dev	727	1122	303	670	149	286	119	16	40
	Test	674	1037	317	607	113	258	104	16	37
Twitter-2017	Train	1745	3562	1508	1638	416	1159	733	15	39
	Dev	577	1176	515	517	144	375	242	16	31
	Test	587	1234	493	573	168	399	263	15	38

Table 1: Statistics of datasets (#S, #A, #Pos, #Neu, #Neg, MA, MS, Mean and Max denote numbers of sentences, aspects, positive aspects, neural aspects, positive aspects, multi aspects in each sentence, multi sentiments in each sentence, mean length and max length).

used as the key and the value,

$$Att_u(z_T^{(L)}, z_V^{(L)}, z_V^{(L)}) = \text{softmax}\left(\frac{(W_Q^u z_T^{(L)})^\top (W_K^u z_V^{(L)})}{\sqrt{d_h/u}}\right)(W_V^u z_V^{(L)}) \tag{8}$$

where $W_Q^u \in \Re^{d_h/u \times N}$ and $\{W_K^u, W_V^u\} \in \Re^{d_h/u \times M}$ are matrices of the query, key and value. With multi-head cross-attention, the final representation of vision-aware text extraction $\bar{p} = [p_1, p_2, .., p_N]$ can be formulated as,

$$\bar{p} = W^p[Att_1, Att_2, ..., Att_U]^\top \tag{9}$$

where $W^p \in \Re^{d_h \times d_h}$ refers to the weight matrix for the multi-head cross-attention.

By passing a MLP and two-layer normalization with two residual connections, the resulting representation is $\hat{p} = [\hat{p}_1, \hat{p}_2, ..., \hat{p}_N]$. To ensure the consistency of representation size, the first residual added the text-only representation.

Different from the text-only tasks, the output layer is a CRF to predict layer sequence y as follows,

$$P(\tilde{\mathbf{y}}|\mathbf{x}) = \frac{\exp(score(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathbf{Y_x}} \exp(score(\mathbf{x}, \mathbf{y}'))} \tag{10}$$

$$score(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^{N} A_{y_n, y_{n+1}} + \sum_{n=0}^{N} w^{y_n} \hat{p}_n \tag{11}$$

where $\mathbf{A}$ is a transition matrix, and its element $A_{i,j}$ represents the score of a transition from tag i to tag j, $w^{y_n} \in \Re^{2 \times d_h}$ is the weights. The loss function is the negative log-probability of the ground truth label,

$$\mathcal{L}^{CM} = -\frac{1}{J}\sum_{j=1}^{J}\left(s(\mathbf{x}^{(j)}, \mathbf{y}^{(j)}) - \underset{\mathbf{y}' \in \mathbf{Y_x^{(j)}}}{\text{logadd}} \exp(s(\mathbf{x}^{(j)}, \mathbf{y}'^{(j)}))\right) \tag{12}$$

Token-Patch Alignment. For matching tokens and patches, there are no annotated labels to supervise the training. Thus, we propose minimizing the Wasserstein distance, also called the earth mover distance (EMD), a measure of the distance between two probability distributions, as shown in Figure 3. Regarding the distribution as a certain amount of earth, the EMD is the minimum cost of turning one pile into another; where the cost is assumed to be the amount of dirt moved times the distance by which it is moved. Based on this, the hidden representation of both text and image for the j-th sample can be assigned with a moving weight,

$$\mathbf{t}^{(j)} = [(\hat{t}_1^{(j)}, w_1^{\mathbf{t}}), (\hat{t}_2^{(j)}, w_2^{\mathbf{t}}), ..., (\hat{t}_N^{(j)}, w_N^{\mathbf{t}})] \tag{13}$$

$$\mathbf{v}^{(j)} = [(\hat{v}_1^{(j)}, w_1^{\mathbf{v}}), (\hat{v}_2^{(j)}, w_2^{\mathbf{v}}), ..., (\hat{v}_M^{(j)}, w_M^{\mathbf{v}})] \tag{14}$$

where $w_n^{\mathbf{t}}$ and $w_m^{\mathbf{v}}$ denote the moving weight, respectively initialized as $1/N$ and $1/M$. The cost of moving $\hat{t}_n$ to $\hat{v}_m$ is a normalized mean squared error (MSE), denoted as,

$$\delta_{m,n} = \text{MSE}(\hat{t}_n, \hat{v}_m) = \frac{1}{d_h}\sum_{d_h}\left\|\frac{\hat{t}_n}{||\hat{t}_n||_2^2} - \frac{\hat{v}_m}{||\hat{v}_m||_2^2}\right\|_2^2 \tag{15}$$

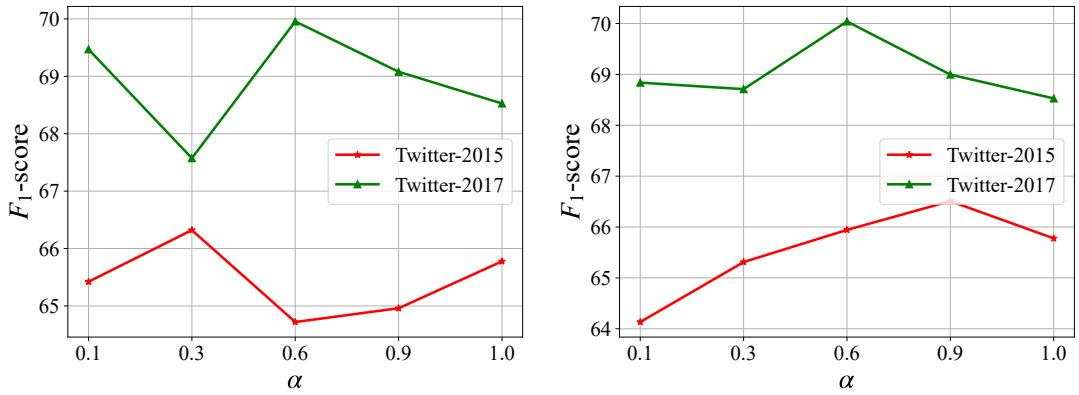
According to Rubner et al. (2000), the target of the token-patch alignment is to find a transfer flow $\mathbf{F}$ that maps the features from $\hat{t}_n$ to $\hat{v}_m$ by minimizing the cumulative cost, defined as,

$$\text{WORK}(\hat{t}_n, \hat{v}_m, \mathbf{F}) = \sum_{n=1}^{N}\sum_{m=1}^{M} f_{m,n}\delta_{m,n} \tag{16}$$

$$\text{s.}t. \quad f_{m,n} \geq 0 \tag{17}$$

$$\sum_{n=1}^{N} f_{m,n} \leq w_n^{\mathbf{t}} \tag{18}$$

$$\sum_{m=1}^{M} f_{m,n} \leq w_m^{\mathbf{v}} \tag{19}$$



(a) Balance coefficient α ($\beta = 1$)

(b) Balance coefficient β ($\alpha = 1$)

Figure 4: Hyper-parameters fine-tuning on different datasets.

$$\sum_{n=1}^{N}\sum_{m=1}^{M} f_{m,n} = \min\left(\sum_{n=1}^{N} w_n^{\mathbf{v}}, \sum_{m=1}^{M} w_m^{\mathbf{v}}\right) \quad (20)$$

where $1 \leq n \leq M$ and $1 \leq m \leq M$ respectively denote the indices of the tokens and image patches. Here, Eq. (17) ensures there is no negative value to impact the result. Eqs. (17) and (18) limit that the number of features which can be sent and received were less than their weights. Eq. (19) ensures the maximum number of features possible are moved. The optimal problem can be solved by the optimal transportation problem, and the cost of token-patch alignment is then defined as the work normalized by the total flow,

$$\mathcal{L}^{WD} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} f_{m,n}\delta_{m,n}}{\sum_{n=1}^{N}\sum_{m=1}^{M} f_{m,n}} \quad (21)$$

2.3 Joint Training

The final objective is a combination over the main task and two auxiliary tasks as follows,

$$\mathcal{L} = \mathcal{L}^{CM} + \alpha\mathcal{L}^{TO} + \beta\mathcal{L}^{WD} \quad (22)$$

where α and β are tradeoff hyper-parameters to control the contribution of each task. For inference, the output of vision-aware text extraction was applied as the results.

3 Experiments

3.1 Datasets and Evaluation Metrics

To evaluate the performance of the dual-encoder transformer with cross-modal alignment, two MABSA benchmark datasets are used, mainly consisting of reviews on Twitter. These datasets are Twitter-2015 and Twitter-2017, originally provided by Zhang et al. (2018) for multimodal named entity recognition and annotated with the sentiment polarity for each aspect by Lu et al. (2018). Table 1 summarizes the statistical characteristics of these two datasets.

Precision, recall, and micro F_1-score are used as evaluation metrics for MABSA. An aspect is regarded as correctly predicted only if the aspect term and polarity respectively match the ground-truth aspect term and corresponding polarity.

3.2 Implementation Details

To evaluate the proposed DTCA model, several baseline models are implemented for comparison, including text-based methods and multimodal methods.

1) Textual methods

- **SPAN** (Hu et al., 2019) is a span-based extract-then-classify framework, where targets are directly extracted from the sentence under the supervision of target span boundaries.
- **D-GCN** (Chen et al., 2020) is a directional graph convolutional network to jointly perform aspect extraction and sentiment analysis with encoding syntactic information.
- **RoBERTa** (Liu et al., 2019) is a pretrained transformer-based model, used as text encoder in the proposed DTCA model.

2) Multimodal methods

Modality	Approaches	Twitter-2015			Twitter-2017		
		F	P	R	F	P	R
Text	SPAN	53.8	53.7	53.9	60.6	59.6	61.7
	D-GCN	59.4	58.3	58.8	64.1	64.2	64.1
	RoBERTa	63.3	62.9	63.7	65.6	65.1	66.2
Text+ Image	UMT-collapse	59.8	58.4	61.3	62.4	62.3	62.4
	OSCGA-collapse	62.5	61.7	63.4	63.7	63.4	64.0
	JML	64.1	65.0	63.2	66.0	66.5	65.5
	DTCA	**68.4**	**67.3**	**69.5**	**70.4**	**69.6**	**71.2**

Table 2: The results of the DTCA model and other models with comparison.

- **UMT-collapse** (Yu et al., 2020) is a directional graph convolutional network used to jointly perform aspect extraction and sentiment analysis with encoding syntactic information.
- **OSCGA-collapse** (Wu et al., 2020) combines object-level image information and character-level text information to predict entities.
- **JML** (Ju et al., 2021) uses a hierarchical framework to bridge the multi-modal connection between MATE and MASC with an auxiliary text-image relation module to ensure the proper exploitation of visual information.

The hyperparameters of all models were finetuned using a grid-search strategy according to the performance on the development set. The hidden size d_h is 768 for both RoBERTa and ViT model. The number of heads in cross-modal self-attention is 8. AdamW optimizer (Loshchilov and Hutter, 2019) with a base learning rate of 2e-5 and warmup decay of 0.1 was used to update all trainable parameters. The maximum length and batch size were respectively set to 60 and 4. For training epochs, we leveraged an early stopping strategy with a patience of 3 to avoid overfitting.

3.3 Hyper-parameters Finetuning

The tradeoff hyper-parameters α and β may impact the final performance of the proposed DTCA method for MABSA. Figure 4 shows the optimal settings according to the final performance on the dev set. We successively fine-tuned each parameter in turn by fixing the other to 1. For both α and β, we used a candidate set of {0.1, 0.3, 0.6, 0.9, 1.0}.

The performance of the proposed DTCA model is optimized when α and β are respectively 0.6 and 0.6 on the **Twitter-2015** dataset and 0.3 and 0.9 on the **Twitter-2017** dataset, the performance of the proposed DTCA model is the best. The results indicate that the use of appropriate parameters can improve the performance.

Model	Twitter-2015			Twitter-2017		
	F1	P	R	F1	P	R
DTCA	67.8	66.9	68.7	70.0	69.5	70.6
w/o TE	67.0	65.9	68.2	68.8	68.6	69.0
w/o TPA	66.5	64.1	68.4	69.1	68.7	69.5
w/o Both	65.6	65.3	65.9	68.7	68.4	69.0

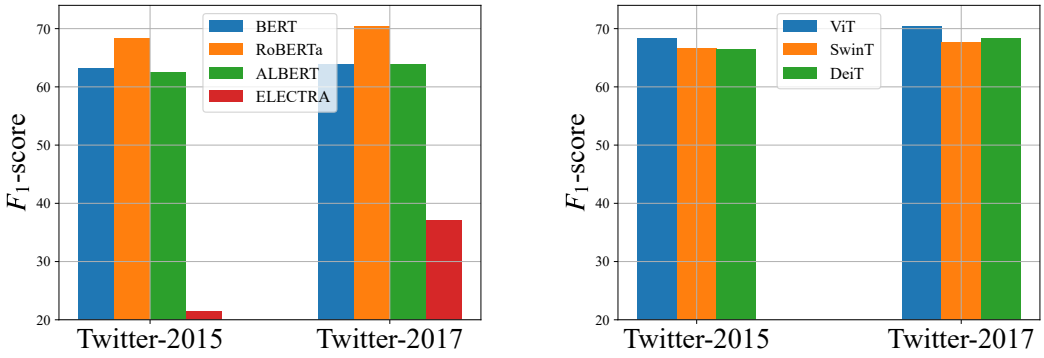
Table 3: The result of ablation. TE: text-only extraction, TPA: token-patch alignment.

3.4 Comparative Results

Table 2 summarizes the comparative results of the proposed DTCA model against several previous methods in terms of precision (P), recall (R), and F_1-score. As indicated, the proposed model outperforms all the baseline models. Compared with the multi-modal baseline with the best performance, i.e. JML, DTCA still shows absolute F_1-score increases of 6.71% and 6.67%. Compared with text-based models, DTCA provides far better results. The F_1-score of the DTCA model on the test set outperforms RoBERTa by 8.06% and 7.32% respectively on **Twitter-2015** and **Twitter-2017**. This indicates that vision-aware text extraction can enable the proposed DTCA model to learn an appropriate representation for MABSA.

3.5 Ablation Study

Table 3 shows the results of an ablation study to further demonstrate the effectiveness of the two auxiliary subtasks, i.e., text-only extraction (TE) and token-patch alignment (TPA). By doing so, we remove TE (w/o TE) and set hyperparameter β as 1.0. Then, we remove TPA (w/o TPA) and set α as 1.0. As indicated, the removal of either one or both subtasks (w/o Both) produce varying degrees of performance decline, indicating that both text-only



(a) Text-encoders (b) Image-encoders

Figure 5: Two results of different modality encoders.

Golden	(a) (*Chris Sale* , Pos)		(b) (*Lebron James* , Neu)	
Visual Modality				
Textual Modality	*Chris Sale records another strikeout , but he ' s only at four in the 7th inning*		*RT @ AndOneNBA : Lebron James on an outlet pass*	
RoBERTa	(*Chris Sale* , Pos)	✓	(*Lebron* , Neu)	×
JML	(*Chris Sale* , Neu)	×	(*Lebron James* , Neu)	✓
DTCA	(*Chris Sale* , Pos)	✓	(*Lebron James* , Neu)	✓

Figure 6: Two examples of the predictions by RoBERTa, JML, DTCA. Pos: Positive, Neu: Neural, Neg: Negative.

extraction and token-patch alignment play indispensable roles in performance improvement.

3.6 Effect of Different Encoder

To investigate the effect of using different encoders, Figure 5 shows the performance of different transform-based encoders for the DTCA model. BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020) and ELECTRA (Clark et al., 2020) were applied as text encoder, while ViT (Dosovitskiy et al., 2021), Swin-Transformer (Liu et al., 2021) and DeiT (Touvron et al., 2021) were applied as image encoder. As shown, RoBERTa achieved the best performance for language modality. For vision modality, the performance margins between different encoders were not obvious, indicating that the text contains enough features to identify the aspect-sentiment pairs, whereas the image sometimes fails to provide complementary information and may even induce noise.

3.7 Case Study

Figure 6 shows a case study of two randomly selected examples. For comparison, both text-only RoBERTa and JML were introduced as baselines. For example (a), although JML can accurately predict the correct aspect term *Chris Sale*, the sentiment of the *Chris Sale* aspect was wrongly predicted. The main reason is the misleading influence of the image. For example (b), RoBERTa only predicts some aspect terms correctly because of the lack of the image relation. In contrast, DTCA can obtain all correct aspect terms and aspect-related sentiment using cross-modal alignment between text and image.

4 Conclusion

This work proposes a dual-encoder transformer with cross-modal alignment for encoding text-image features into the representations for MABSA tasks. A multitask learning architecture containing three subtasks was applied to integrate both text and image modalities. In addition to the co-attention module, the token-patch alignment was introduced to improve model training effectiveness. Empirical experiments show the model improved the performance for MABSA in the Twitter-2015 and Twitter-2017 datasets. In addition, ablation and case studies further indicate the effectiveness of the proposed model.

Future work will extend the proposed method to more multi-modal tasks, such as multi-modal

MRC, ASTE and dialogue.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051, and the Ministry of Science and Technology, Taiwan, ROC, under Grant No.MOST 111-2628-E-155-001-MY2. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING-2020)*, pages 272–279.

Kevin Clark, Minh-Thang Luong, and Quoc V Le. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR-2020)*, pages 756–773.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR-2021)*, pages 381–401.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV-2015)*, pages 1440–1448.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2016)*, pages 770–778.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. *arXiv preprint arXiv:2204.05356*.

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 the 57th Annual Meeting of the Association for Computational Linguistics (ACL-2019)*, 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 the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021)*, pages 4395–4405.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite bert for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR-2020)*, pages 1034–1050.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the 16th European Conference on Computer Vision (ECCV-2020)*, pages 121–137.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL-2022)*, pages 2149–2159.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV-2021)*, pages 9992–10002.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR-2019)*, pages 433–451.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL-2018)*, pages 1990–1999.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR-2015)*, pages 349–362.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML-2021)*, volume 139, pages 10347–10357.

Milad Vazan and Jafar Razmara. 2021. Jointly modeling aspect and polarity for aspect-based sentiment analysis in persian reviews. *arXiv preprint arXiv:2109.07680*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated concatenation of embeddings for structured prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2021)*, pages 2643–2660.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM-Multimedia-2020)*, pages 1038–1046.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. MAF: A general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (ACM-WSDM-2022)*, pages 1215–1223.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS-2019)*, pages 5754–5764.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020)*, pages 3342–3352.

Jianfei Yu, Jieming Wang, Rui Xia, and Junjie Li. 2022. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence(IJCAI-2022)*, pages 4482–4488.

Li Yuan, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2022. Hierarchical template transformer for fine-grained sentiment controllable generation. *Information Processing & Management*, 59(5):103048.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-2018)*, pages 5674–5681.

You Zhang, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2021a. MA-BERT: Learning representation by incorporating multi-attribute knowledge in transformers. In *Findings of the Association for Computational Linguistics(ACL-IJCNLP-2021)*, pages 2338–2343.

You Zhang, Jin Wang, and Xuejie Zhang. 2021b. Learning sentiment sentence representation with multiview attention model. *Information Sciences*, 571:459–474.