Multi-Modality Cross Attention Network for Image and Sentence Matching

Xi Wei[1], Tianzhu Zhang[1,*], Yan Li[2], Yongdong Zhang[1], Feng Wu[1]
[1] University of Science and Technology of China; [2] Kuaishou Technology
wx33921@mail.ustc.edu.cn; {tzzhang,fengwu,zhyd73}@ustc.edu.cn; liyan@kuaishou.com

Abstract

The key of image and sentence matching is to accurately measure the visual-semantic similarity between an image and a sentence. However, most existing methods make use of only the intra-modality relationship within each modality or the inter-modality relationship between image regions and sentence words for the cross-modal matching task. Different from them, in this work, we propose a novel Multi-Modality Cross Attention (MMCA) Network for image and sentence matching by jointly modeling the intra-modality and inter-modality relationships of image regions and sentence words in a unified deep model. In the proposed MMCA, we design a novel cross-attention mechanism, which is able to exploit not only the intra-modality relationship within each modality, but also the inter-modality relationship between image regions and sentence words to complement and enhance each other for image and sentence matching. Extensive experimental results on two standard benchmarks including Flickr30K and MS-COCO demonstrate that the proposed model performs favorably against state-of-the-art image and sentence matching methods.

1. Introduction

Image and sentence matching is one of the fundamental tasks in the field of vision and language [26, 6, 13]. The goal of such a cross-modal matching task is how to accurately measure the visual-semantic similarity between an image and a sentence, and is related to many vision-language tasks including image-sentence cross-modal retrieval [21, 48, 55], visual captioning [1, 5], visual grounding [8, 53] and visual question answering [1, 52, 25, 25]. This task has drawn remarkable attention and has been widely adopted to various applications [12, 43, 13, 48], e.g., finding similar sentences given an image query for image annotation and caption, and retrieving matched images with a sentence query for image search. Although significant progress has been achieved in recent years, it is still a challenging problem because it requires the understanding
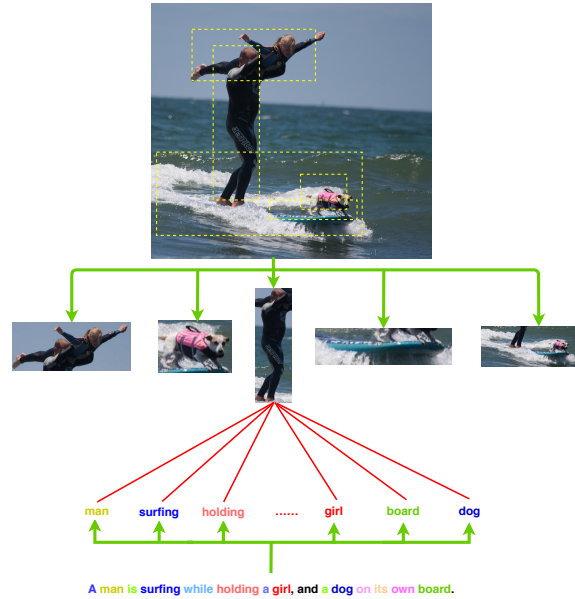
---

*Corresponding Author

Figure 1. Fine-grained representation for visual and textual elements. Our method can jointly model both intra-modality and inter-modality relationships in a deep network.

of language semantics, visual contents, and cross-modal relationships and alignments [38].

Due to the huge visual-semantic discrepancy between vision and language [6, 42], matching images and sentences is still far from being solved. Recently, various methods have been proposed for this problem, which can be classified into two categories including one-to-one matching [6, 19, 28, 50, 10] and many-to-many matching [14, 12, 21, 26, 48]. One-to-one matching methods usually extract global representations for image and sentence, and then associate them by exploiting visual-semantic embedding [42]. Most previous methods embed images and sentences independently into the same embedding space, and then measure their similarities by feature distances in the joint space. Driven by the success of deep learning, the main stream has been changed to modality-specific deep feature learning, e.g., learning CNN for image and RNN for sentence. For visual-content understanding, several deep backbone models have been developed including VGG, ResNet, GoogleNet [11, 34, 37],

and demonstrated their effectiveness on large vision datasets [3, 20]. In terms of language understanding, several methods [4, 31, 39] have been proposed towards building a universal backbone model with large-scale contextualized language model pre-training [39, 4], which has improved performances on various tasks to significant levels [31, 41]. For the cross-modality association, the simplest way is to learn projection functions to map visual and textual data into the same embedding space, such as, canonical correlation objective [19, 50], structured objective [7, 26, 9, 52, 40]. However, such independent embedding approaches ignore the fact that the global similarity commonly arises from a complex aggregation of local similarities between image-sentence fragments (objects in an image and words in a sentence) [12]. As a result, most of existing methods might lead to suboptimal features for image-sentence matching.

To deal with the above issues, many-to-many matching methods have been proposed to take the relationships between regions of the image and words of the sentence into consideration [12, 14, 15, 26, 21, 48]. Most existing methods compare many pairs of image regions and sentence words, and aggregate their local similarities [14, 15, 18, 35]. Generally, incorporating the relationships between image regions and sentence words could benefit in capturing the fine-grained cross-modal cues for image and sentence matching. To achieve this goal, various methods have been proposed [15, 13, 14, 29, 52, 21, 48, 12], which can be roughly grouped into two categories including inter-modality based methods [14, 21, 12, 26, 13] and intra-modality based methods [48, 38]. The inter-modality based methods [18, 12, 26, 21] mainly focus on discovering possible relationships between image regions and sentence words, which make great progress in considering the interactions between regions and words. As shown in Figure 1, if the word 'man' shares the inter-modality information with the corresponding regions in the image, it would be easier to capture the relevance among these two heterogeneous data. However, most of existing methods ignore the connections among vision-vision elements or language-language elements. The intra-modality based methods stress the relations within each modality for image regions or sentence words [48], which ignores the inter-modality relations across different modalities. As shown in Figure 1, if the word 'man' has a tight connection with the word 'surfing', 'holding', 'girl' in the sentence, it would have a better representation to help obtain the global feature for the whole sentence. Based on the above discussions, to the best of our knowledge, the inter-modality and intra-modality relationships are not jointly investigated in a unified framework for solving the image and sentence matching problem. As shown in Figure 1, the intra-modality relationship within each modality and the inter-modality relationship between image regions and sentence words can complement and enhance each other for image and sentence matching.

Motivated by the above discussions, we propose a novel Multi-Modality Cross Attention Network for image and sentence matching by jointly modeling inter-modality relationship and intra-modality relationship of image regions and sentence words in a unified deep model. To achieve a robust cross-modal matching, we design two effective attention modules including self-attention module and cross-attention module, which play important roles in modeling the relationships of intra-modality and inter-modality. In the self-attention module, we employ the bottom-up model [1] to extract features of salient image regions. Meanwhile we use the word token embeddings as the language elements. Then we independently feed the image regions into the Transformer [39] unit and the word tokens into BERT model [4] to discover the intra-modality relationship. Then, we can obtain the global representation by aggregating these fragments features. In the cross-attention module, we stack the representations of image regions and sentence words and then pass them into another Transformer unit followed by a 1d-CNN [16] and a pooling operation to fuse both inter-modality and intra-modality information. Then based on the updated features for visual and textual data, we can predict the similarity score of the input image and sentence.

The major contributions of this work can be summarized as follows. (1) We proposed a novel Multi-Modality Cross Attention Network for image and sentence matching by jointly modeling intra-modality relationship and inter-modality relationship of image regions and sentence words in a unified deep model. (2) To achieve a robust cross-modal matching, we propose a novel cross-attention module, which is able to exploit not only the intra-modality relationship within each modality, but also the inter-modality relationship between image regions and sentence words to complement and enhance each other for image and sentence matching. (3) Extensive experimental results on two standard benchmarks including Flickr30K and MS-COCO demonstrate that the proposed model performs favorably against state-of-the-art image and sentence matching methods.

2. Related Work

In this section, we discuss related works about image and sentence matching. Following [12], we roughly divide the related methods into two categories including one-to-one matching and many-to-many matching. Furthermore, we briefly review attention mechanism based methods.

One-to-one Matching. A rich line of early studies extract global representations for image and sentence, and then associate them with the hinge-based triplet ranking loss in which the matched image-sentence pairs have small distances [18, 19, 28, 50]. In [6], Faghri et al. attempt to use hard negative mining in the triplet loss function and achieve
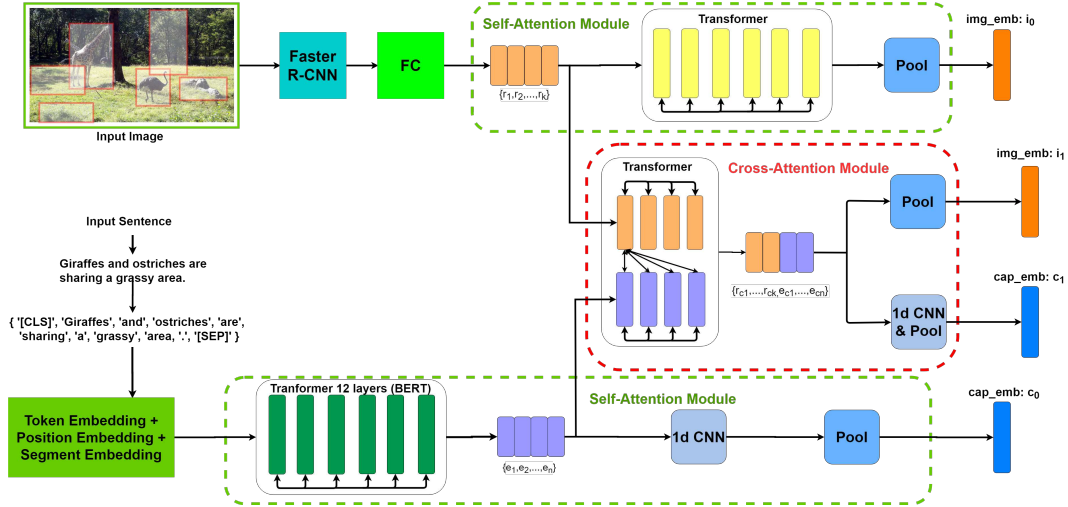
Figure 2. Our proposed Multi-Modality Cross Attention Network, consisting of the self-attention module and the cross-attention module. The self-attention module is exhibited in the green dashed blocks, while the cross-attention module is shown in the red dashed block. For more details, please refer to the text.

a significant improvement. In [10] and [28], the generative objectives are combined with the cross-view feature embedding learning to learn more discriminative representations for visual and textual data. Meanwhile, Yan et al. [50] associate features of image and sentence with deep canonical correlation analysis where the true matched image-sentence pairs have a high correlation. With a similar objective, Klein et al. [19] take use of Fisher Vectors (FV) to obtain a discriminative sentence representation. Furthermore, Lev et al. [22] exploit RNN to encode FV leading to better performance. However, the above methods ignore the fact that the global similarity arises from a complex aggregation of the latent vision-language correspondences at the level of image regions and sentence words.

Many-to-many Matching. In the field of vision and language, it is more and more popular to consider the fine-grained alignments between image regions and sentence words. In [14], it is the first work to perform local similarity learning between fragments of image regions and sentence words with a structured objective. In [12], a selective multi-modal long short term memory network is proposed for the instance-aware image and sentence matching. Similarly, in [26], a dual attentional network is proposed to capture the fine-grained interplay between vision and language through multiple steps. However, this work takes a multi-step approach to realize the feature alignment between the whole image and sentence, which is less interpretable. S-CAN [21] is proposed by using a stacked cross attention mechanism to discover all the alignments between salient objects and words. But it fails to consider the relationships within image regions or sentence words. And later, SAEM [48] resorts to a self-attention mechanism to explore the relationship within each modality, yet it ignores the relationship across different modalities. However, few methods have been proposed to investigate the inter-modality and

intra-modality relationships jointly in a unified framework for image and sentence matching.

Attention Based Methods. Attention mechanism has been developed to simulate the human behavior that humans selectively use part of the data to make a decision. It has been widely applied to various visual and textual tasks, including image classification [53], object detection [47], image captioning [49], sentiment classification [45], neural machine translation [24, 39], sentence summarization [33], etc. Recently, the attention mechanism has also been applied to the cross-modality matching task. In [26], the dual attentional network is proposed to align different visual regions and words in sentences by multiple steps. Ba et al. [2] present a recurrent attention model that can attend to some label relevant image regions for object recognition. In [21], a stacked cross attention mechanism is adopted to discover the latent alignments using both image regions and words in a sentence as context, but ignores the intra-modality relationship. Inspired by the Transformer in machine translation [39], lots of recent works [52, 38, 48, 9] take use of the Transformer model to implement the self-attention mechanism. However, they mainly explore the intra-modality relationships. Different from existing methods, the proposed cross-attention model can discover both the intra-modality and inter-modality relationships jointly for image and sentence matching in a unified model.

3. Multi-Modality Cross Attention Network

3.1. Overview

As shown in Figure 2, our Multi-Modality Cross Attention Network mainly consists of two modules, the self-attention module and the cross-attention module, demonstrated in the green dashed blocks and red dash dashed block in Figure 2, respectively. Given an image and sentence pair, we first feed the image into the bottom-up atten-

tion model [32] pre-trained on Visual Genome [20] to extract features for image regions. Meanwhile, we use Word-Piece tokens of each sentence as the fragments in the textual modality. Based on these extracted fine-grained representations for image regions and sentence words, we model the intra-modality relationship with the Self-Attention Module, and adopt the Cross-Attention Module to model the inter-modality and intra-modality relationships for image regions and sentence words. By considering both intra-modality and inter-modality relationships into consideration, the features discriminative ability of image and sentence fragments can be improved. Then the 1d-CNN and pool operation are used to aggregate these fragment representations, resembling bag of visual words model which has shown success in the content based image indexing and retrieval [30] in early ages. As shown in Figure 2, we get two pairs of embeddings for the given image-sentence pair (i_0, c_0) and (i_1, c_1), which are used for image and sentence matching. In the training stage, we construct the bi-directional triplet loss with the hard negative mining to optimize the parameters in our model. Details are introduced as follows.

3.2. Instance Candidate Extraction

Image Instance Candidates. Given an image I, we use the bottom-up attention model [1] pre-trained on Visual Genome [20] to extract region features. The output is a set of region features $\mathbf{O} = \{o_1, o_2, ..., o_k\}$, where each o_i is defined as the mean-pooled convolutional feature for the ith region. The pretrained model is fixed during training. And we add a fully-connect layer to transform the region features to fit our task. We denote the transformed feature as $\mathbf{R} = \{r_1, r_2, ..., r_k\}$, with r_i corresponding to the transformed feature of o_i.

Sentence Instance Candidates. Following [4], which has made great progress in machine translation, we use Word-Piece tokens of sentence T as the fragments in textual modality. And the final embedding for every word is the combination of its' token embedding, position embedding and segment embedding, denoted as $\mathbf{X} = \{x_1, x_2, ..., x_n\}$.

These image region features and word embeddings are further fed into our Multi-Modality Cross Attention Network to fuse both the intra-modality and inter-modality information.

3.3. Self-Attention Module

In this section, we introduce how to utilize the self-attention module to model the intra-modality relationship for image regions and sentence words, respectively. We first give a glance at the paradigm of the attention function.

An attention module can be described as mapping a query and a set of key-value pairs to an output. The output of attention function is a weighted sum of the value, where the weight matrix, or affinity matrix, is determined by query and its corresponding key. Specifically, for self-attention mechanism, queries, keys and values are equal.
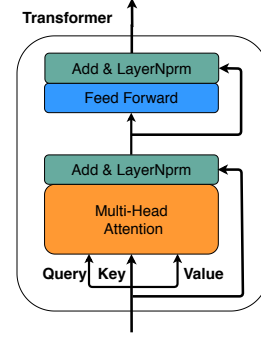

Figure 3. The Transformer unit with the multi-head sub-layer and the position wise feed-forward sub-layer. Meanwhile, residual connections followed by the layer normalization are also applied around each of the two sub-layers.

Following the philosophy of [39], we apply the Transformer to implement the attention function. As shown in Figure 3, the Transformer consists of two sub-layers, the multi-head self-attention sub-layer, and the position wise feed-forward sub-layer.

In the multi-head self-attention sub-layer, the attention is calculated h times, making it to be multi-headed. This is achieved by projecting the queries ($\mathbf{Q}$), keys ($\mathbf{K}$) and values ($\mathbf{V}$) with h times by using different learnable linear projections. To be specific, given a set of fragments $\mathbf{F} = \{f_1; f_2; ...; f_k\}$, where $f_i \in R^{1 \times d_f}$ and $\mathbf{F} \in R^{k \times d_f}$ (indicating the stacked features of image regions or sentence words), we firstly calculate the query, key and value for the input: $\mathbf{Q}_F = \mathbf{F}\mathbf{W}_i^Q$, $\mathbf{K}_F = \mathbf{F}\mathbf{W}_i^K$, $\mathbf{V}_F = \mathbf{F}\mathbf{W}_i^V$, where $\mathbf{W}_i^Q \in R^{d_f \times d_k}$, $\mathbf{W}_i^K \in R^{d_f \times d_k}$, $\mathbf{W}_i^V \in R^{d_f \times d_v}$, the subscript i donates for the i-th head. Then we can obtain the weight matrix or affinity matrix with 'Scaled Dot-Product Attention'. Furthermore, the weighted sum of the value is computed through the following equation:

$$Attention\left(\mathbf{Q}_F, \mathbf{K}_F, \mathbf{V}_F\right) = softmax\left(\frac{\mathbf{Q}_F \mathbf{K}_F^T}{\sqrt{d_k}}\right)\mathbf{V}_F.$$
(1)

After that, we compute the values for all heads and concatenate them together with equations:

$$head_i = Attention\left(\mathbf{F}\mathbf{W}_i^Q, \mathbf{F}\mathbf{W}_i^K, \mathbf{F}\mathbf{W}_i^V\right),$$
(2)

$$MultiHead\left(\mathbf{F}\right) = concat\left(head_1, ..., head_h\right)\mathbf{W}^O,$$
(3)

where $\mathbf{W}^O \in R^{h d_v \times d_k}$, and h is the number of the heads.

Aiming to further adjust the fragment representations, the position wise feed-forward sub-layer transforms each fragment separately and identically with two fully-connected layers. And it can be described as:

$$FFN\left(x\right) = ReLu\left(x\mathbf{W}_1 + b_1\right)\mathbf{W}_2 + b_2,$$
(4)

where $x \in R^{1 \times d_x}$, $\mathbf{W}_1 \in R^{d_x \times d_x}$, $\mathbf{W}_2 \in R^{d_x \times d_x}$, $b_1 \in R^{1 \times d_x}$ and $b_2 \in R^{1 \times d_x}$. Furthermore, the residual connections [11] followed by a layer normalization are also

applied after each of the two sub-layers to facilitate optimization.

With the above self-attention unit, every image region or sentence word can attend to the features of other fragments in the same modality. As shown in the top green dashed block in Figure 2, for image I with the fine-grained representation $\mathbf{R} = \{r_1, r_2, ..., r_k\}$, we adapt the above Transformer unit and produce the features $\mathbf{R}_s = \{r_{s1}, r_{s2}, ..., r_{sk}\}$ containing region-region relations. Next, we aggregate the representations of image regions through a simple but effective average pooling operation:

$$i_0 = \frac{1}{k} \sum_{i=1}^{k} r_{si}. \tag{5}$$

And the ℓ_2 normalization is also applied to adjust the ultimate global representation $i_0 \in R^{d \times 1}$ for this image. As the i_0 aggregates the fragment features in $\mathbf{R}_s$, the representation for image I contains the intra-modality relations.

For the textual data modeling, we feed the tokens ($\mathbf{X} = \{x_1, x_2, ..., x_n\}$) of the sentence T into the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model [4] as shown in the bottom green dashed block in Figure 2. The BERT consists of multiple Transformer units, and its output $\mathbf{E} = \{e_1, e_2, ..., e_n\}$ naturally includes the intra-modality information. Then the 1-dim convolution neural networks [16] are used to extract the local context information. In particularly, three window sizes (uni-gram, bi-gram and tri-gram) are used to capture the phrase level information. The convolutional output using the window size l for the k-th word is:

$$p_{l,k} = ReLU\left(\mathbf{W}_l e_{k:k+l-1} + b_l,\right) \quad, \quad l = 1, 2, 3 \tag{6}$$

where $\mathbf{W}_l$ is the convolution filter matrix, and b_l is the bias. Next, the max-pooling operation across all word locations is carried out: $q_l = max\{p_{l,1}, ..., p_{l,n}\}$. Then we concatenate q_1, q_2, q_3 and pass it into a fully connected layer followed by the ℓ_2 normalization to get the final sentence embedding c_0:

$$c_0 = LayerNorm\left(\mathbf{W}_e \, concat\left(q_1, q_2, q_3\right) + b_e\right), \tag{7}$$

where $W_e \in R^{d \times 3d}$ and $b_e \in R^{d \times 1}$. Similarly, $c_0 \in R^{d \times 1}$ models the intra-modality relationship of textual data.

3.4. Cross-Attention Module

Although the above self-attention module can effectively exploit the intra-modality relationship, the inter-modality relationship, e.g., the relationship of image regions and sentence words is not explored. In this section, we introduce how to model both the inter-modality and intra-modality relationships in a unified model with our Cross-Attention Module. The detailed introduction are as follows.

As shown in the red dashed block in Figure 2, the cross-attention module takes the stacked features of image regions and sentence words $\mathbf{Y} = \binom{\mathbf{R}}{\mathbf{E}} = \{r_1; ...; r_k; e_1; ...; e_n\}$ as the input, where $\mathbf{Y} \in R^{(k+n) \times d_x}$. Then $\mathbf{Y}$ is passed into another Transformer unit. Here, the query, key and value for the fragments are formed with the following equations:

$$\mathbf{K}_Y = \mathbf{Y}\mathbf{W}^K = \begin{pmatrix} \mathbf{R}\mathbf{W}^K \\ \mathbf{E}\mathbf{W}^K \end{pmatrix} = \begin{pmatrix} \mathbf{K}_R \\ \mathbf{K}_E \end{pmatrix}, \tag{8}$$

$$\mathbf{Q}_Y = \mathbf{Y}\mathbf{W}^Q = \begin{pmatrix} \mathbf{R}\mathbf{W}^Q \\ \mathbf{E}\mathbf{W}^Q \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_R \\ \mathbf{Q}_E \end{pmatrix}, \tag{9}$$

$$\mathbf{V}_Y = \mathbf{Y}\mathbf{W}^V = \begin{pmatrix} \mathbf{R}\mathbf{W}^V \\ \mathbf{E}\mathbf{W}^V \end{pmatrix} = \begin{pmatrix} \mathbf{V}_R \\ \mathbf{V}_E \end{pmatrix}. \tag{10}$$

Next, the 'Scaled Dot-Product Attention' is carried out as defined in Eq.(11):

$$Attention\left(\mathbf{Q}_Y, \mathbf{K}_Y, \mathbf{V}_Y\right) = softmax\left(\frac{\mathbf{Q}_Y \mathbf{K}_Y^T}{\sqrt{d}}\right) \mathbf{V}_Y. \tag{11}$$

To keep our derivation simple and easy to be understood, we get rid of the softmax and scaled function in the above equation, which does not affect the core idea of our attention mechanism. And it can be expanded as follows:

$$
\begin{aligned}
\mathbf{Q}_Y \mathbf{K}_Y^T \cdot \mathbf{V}_Y &= \begin{pmatrix} \mathbf{Q}_R \\ \mathbf{Q}_E \end{pmatrix} \begin{pmatrix} \mathbf{K}_R^T & \mathbf{K}_E^T \end{pmatrix} \cdot \begin{pmatrix} \mathbf{V}_R \\ \mathbf{V}_E \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{Q}_R \mathbf{K}_R^T & \mathbf{Q}_R \mathbf{K}_E^T \\ \mathbf{Q}_E \mathbf{K}_R^T & \mathbf{Q}_E \mathbf{K}_E^T \end{pmatrix} \cdot \begin{pmatrix} \mathbf{V}_R \\ \mathbf{V}_E \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{Q}_R \mathbf{K}_R^T \mathbf{V}_R + \mathbf{Q}_R \mathbf{K}_E^T \mathbf{V}_E \\ \mathbf{Q}_E \mathbf{K}_E^T \mathbf{V}_E + \mathbf{Q}_E \mathbf{K}_R^T \mathbf{V}_R \end{pmatrix}.
\end{aligned} \tag{12}
$$

As we know $\begin{pmatrix} \mathbf{R}_{up} \\ \mathbf{E}_{up} \end{pmatrix} = \mathbf{Q}_Y \mathbf{K}_Y^T \cdot \mathbf{V}_Y$, the updated features for visual and textual fragments are:

$$\mathbf{R}_{up} = \{r_{up1}; ...; r_{upk}\} = \mathbf{Q}_R \mathbf{K}_R^T \mathbf{V}_R + \mathbf{Q}_R \mathbf{K}_E^T \mathbf{V}_E, \tag{13}$$

$$\mathbf{E}_{up} = \{e_{up1}; ...; e_{upk}\} = \mathbf{Q}_E \mathbf{K}_E^T \mathbf{V}_E + \mathbf{Q}_E \mathbf{K}_R^T \mathbf{V}_R. \tag{14}$$

This result shows that the output of the multi-head sub-layer in this Transformer unit synchronously takes the inter-modality and intra-modality relationships into consideration. Then $\begin{pmatrix} \mathbf{R}_{up} \\ \mathbf{E}_{up} \end{pmatrix}$ can be send into the followed position wise feed-forward sub-layer. Finally, we get the output of the Transformer unit in the cross-attention module, and write it as: $\mathbf{Y}_c = \begin{pmatrix} \mathbf{R}_c \\ \mathbf{E}_c \end{pmatrix}$.

In order to obtain the final representations for the whole image and sentence, we split $\mathbf{Y}_c$ into $\mathbf{R}_c = \{r_{c1}...r_{ck}\}$ and $\mathbf{E}_c = \{e_{c1}...e_{cn}\}$, and once again, pass them into an average pool layer (for image regions $\mathbf{R}_c$) or an 1d-CNN layer followed by a max pool layer (for words in sentence $\mathbf{E}_c$), which is quite similar to the last few operations in the self-attention module. So we have the final embedding $i_1 \in R^{d \times 1}$ and $c_1 \in R^{d \times 1}$ for image and sentence.

3.5. Alignment Objective

Based on the above discussion, we can learn two pairs of embeddings i.e., (i_0, c_0) and (i_1, c_1) for the given image-sentence pair (I, T). Since the embeddings are scaled to have a unit norm, we define the similarity score for image I and sentence T as the weighted sum of two inner products, i.e., $S(I, T) = i_0 \cdot c_0 + \alpha(i_1 \cdot c_1)$, where α is a hyper-parameter which balances the impact of the self-attention module and the cross-attention module.

Then our model can be trained with a bi-directional triplet ranking loss which encourages the similarity scores of matched images and sentences to be larger than those of mismatched ones.

$$
\begin{aligned}
\mathcal{L} = &\ max \left[0, m - S(I, T) + S\left(I, \hat{T}\right) \right] \\
&+ max \left[0, m - S(I, T) + S\left(\hat{I}, T\right) \right],
\end{aligned} \tag{15}
$$

where m denotes the margin, (I, T) denotes the true matched image-sentence pair, and $\hat{I}$, $\hat{T}$ stand for the hard negatives in a mini-batch, i.e., $\hat{I} = argmax_{x \neq I} S(x, T)$ and $\hat{T} = argmax_{y \neq T} S(I, y)$. In practice, we only use the hard negatives in a mini-batch, instead of summing over all the negative samples, which has proved to be effective for the retrieval performance as in [6].

4. Experimental Results

To demonstrate the effectiveness of our proposed method, we carry out extensive experiments on two public available datasets including MS-COCO [23] and Flickr30K [51]. Conventionally, we take Recall at K (R@K) as the evaluation metric, i.e. the fraction of queries for which the correct item is retrieved in the closest K points to the query. Besides, we conduct ablation studies to thoroughly investigate our method.

4.1. Datesets and Protocols

MS-COCO [23] is one of the most popular dataset for the image and sentence matching task. It contains 123287 images, and each image is annotated with five text descriptions. The average length of captions is 8.7 after a rare word removal. In [14], the dataset is split into 82783 training images, 5,000 validation images and 5000 test images. We follow [6] to add 30504 images that are originally in the validation set of MS-COCO but have been left out in this split into the training set.

Flickr30K [51] consists of 31000 images collected from the Flickr website. And every image contains 5 text descriptions. We take the same split for training, validation and testing set as in [14]. There are 1000 images for validation and 1000 images for testing, and the rest for training.

4.2. Implementation Details

The proposed Multi-Modality Cross-Attention Network is implemented in PyTorch framework [27] with a NVIDIA GeForce GTX 2080Ti GPU.

In the self-attention module, for the image branch, the image region feature vector extracted by a bottom-up attention [1] is 2048-dimensional, and we add a fully-connect layer to transform it to a d-dimensional vector before feeding them into a Transformer unit with 16 heads. As for the textual data in the self-attention module, we use the pre-trained BERT model [4] including 12 self-attention layers, 12 heads, 768 hidden units for each token. For simplicity, the weights of BERT model is fixed during the training stage. In the 1-dim convolution neural networks, we use 256 filters for each filter size. In the cross-attention module, we apply a Transformer unit with 16 heads for implementation.

The model is trained for 20 epochs with the Adam optimizer [17]. We start training with a learning rate 0.0002 for the first 10 epochs, and then decay the learning rate by 0.1 for the rest epochs. The batch-size is set to 64 for all experiments. The margins for the hinge triplet loss is set to 0.2, i.e., $m = 0.2$. Note that since the size of the training set for Flickr30k and MS-COCO is different, the actual number of iterations in each epoch can vary. At last, for evaluation on the test set, we tackle the over-fitting by choosing the snapshot of the model based on the validation set.

4.3. Performance Comparison

We compare with several recent state-of-the-art methods on the Flickr30k [51] and MS-COCO [23] datasets in Table 1, Table 2 and Table 3. We can find that our proposed Multi-Modality Cross Attention Network can achieve much better performance.

Results on Flickr30K. Table 1 presents the quantitative results on Flickr30K where our proposed method outperforms recent approaches in both image to sentence retrieval and sentence to image retrieval, achieving 74.2%, 92.8%, 96.4% for R@1, R@5 and R@10 in image to sentence retrieval task. And the performance on sentence to image retrieval is 54.8%, 81.4%, and 87.85% for R@1, R@5 and R@10, respectively. Besides, almost all the best methods are based on Faster R-CNN [32], which shows that the methods can usually work better if they take the fine-grained image regions for the global image representation. Different from previous works which either neglect the interactions between regions and words, or ignore the textual-textual and visual-visual relationships, our model jointly models both inter-modality and intra-modality relationships in a unified deep model.

Results on MS-COCO Table 2 and Table 3 list the experimental results on MS-COCO (1K testing set and 5K testing set respectively) compared with previous methods. It can be seen from Table 2 that our proposed method outperforms the recent approaches. When measured by R@1, our model outperforms the best baseline by 3.6% and 3.8% on the image-to-text task and the text-to-image task, respectively. On the 5K testing set, our proposed method out-

Table 1. Comparison results of the cross-modal retrieval on the Flickr30K dataset in terms of Recall@K($R@K$).

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA (R-CNN, AlexNet) [14]	22.2	48.2	61.4	15.2	37.7	50.5
DSPE(VGG, Fisher vector) [43]	40.3	68.9	79.9	29.7	60.1	72.1
JGCAR(VGG) [44]	44.9	75.3	82.7	35.2	62.0	72.4
CMPM (ResNet) [54]	49.6	76.8	86.1	37.3	65.7	75.5
DAN (ResNet) [26]	55.0	81.8	89.0	39.4	69.2	79.1
VSE++ (ResNet) [6]	52.9	87.2	-	39.6	-	79.5
DPC (ResNet) [55]	55.6	81.9	89.5	39.1	69.2	80.9
SCO (ResNet) [13]	55.5	82.0	89.3	41.1	70.5	80.1
CAMP (Faster R-CNN, ResNet) [46]	68.1	89.7	95.2	51.5	77.1	85.3
SCAN (Faster R-CNN, ResNet) [21]	67.4	90.3	95.8	48.6	77.7	85.2
SAEM (Faster R-CNN, ResNet) [48]	69.1	91.0	95.1	52.4	81.1	**88.1**
MMCA	**74.2**	**92.8**	**96.4**	**54.8**	**81.4**	87.8

Table 2. Comparison results of the cross-modal retrieval on the MS-COCO 1K testing set in terms of Recall@K(R@K).

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA (R-CNN, AlexNet) [14]	38.4	69.9	80.5	27.4	60.2	74.8
DSPE(VGG, Fisher vector) [43]	50.1	79.7	89.2	39.6	75.2	86.9
CMPM (ResNet) [54]	56.1	86.3	92.9	44.6	78.8	89.0
JGCAR(VGG) [44]	52.7	82.6	90.5	40.2	74.8	85.7
VSE++ (ResNet) [6]	64.6	-	95.7	52.0	-	92.0
DPC (ResNet) [55]	65.6	89.8	95.5	47.1	79.9	90.0
GXN (ResNet) [10]	68.5	-	97.9	56.6	-	94.5
PVSE (ResNet) [36]	69.2	91.6	96.6	55.2	86.5	93.7
SCO (ResNet) [13]	69.9	92.9	97.5	56.7	87.5	94.8
CMPM (ResNet) [54]	56.1	86.3	92.9	44.6	78.8	89.0
CAMP (Faster R-CNN, ResNet) [46]	72.3	94.8	98.3	58.5	87.9	95.0
SCAN (Faster R-CNN, ResNet) [21]	72.7	94.8	98.4	58.8	88.4	94.8
SAEM (Faster R-CNN, ResNet) [48]	71.2	94.1	**97.7**	57.8	88.6	94.9
MMCA	**74.8**	**95.6**	**97.7**	**61.6**	**89.8**	**95.2**

performs recent approaches, achieving 54.0% for R@1 in the image to sentence retrieval task and 38.7% for R@1 in the sentence to image retrieval task. The superiority of our model can be attributed to its ability to exploit region-region, word-word and region-word relationships through the self-attention module and the cross-attention module a unified network. Thus we can obtain more suitable embeddings for measuring the relevance between visual and textual data, and get the retrieval task better.

4.4. Ablation Studies and Analysis

First of all, we conduct ablation studies on Flickr30K and MS-COCO 1K testing set to revisit the effect of the dimensionality of the hidden space. The results for the image-sentence retrieval with varying dimensions are shown in Table 4 and Table 5. And it can be seen from the table that the performance of our model first increases and then decreases, with the increasing of the hidden space dimension. We get the best result when the dimension of hidden space is set to 256 on both Flickr30K and MS-COCO datasets. The result indicates that larger dimensions do not always lead to better performance. And it may be because that larger dimensions make the model difficult to train. So it is necessary to choose an appropriate middle-sized dimensionality for our model.

Secondly, we test the effect of different values of the hyper-parameter α in equation $S(I,T) = i_0 \cdot c_0 + \alpha(i_1 \cdot c_1)$. As we can see, α acts as a balancer to control the impact of the self-attention module and the cross-attention module for the final matching score. The experimental results are shown in Table 6. We obtain the best performance when α is set to 0.2. Essentially, if α equals to zero, the model only takes the intra-modality relationships into consideration, which leads to drop on the performance. Moreover, α with large values also has a negative impact on the result, which may be because that the balance of the intra-modality and inter-modality relationships should be considered for the image and sentence matching task. When α is too large, we may loss the essential information for visual contents and language semantics.

Table 3. Comparison results of the cross-modal retrieval on the MS-COCO 5K testing set in terms of Recall@K(R@K).

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
VSE++ (ResNet) [6]	41.3	-	81.2	30.3	-	72.4
DPC (ResNet) [55]	41.2	70.5	81.1	25.3	53.4	66.4
GXN (ResNet) [10]	42.0	-	84.7	31.7	-	74.6
SCO (ResNet) [13]	42.8	72.3	83.0	33.1	62.9	75.5
PVSE (ResNet) [36]	45.2	74.2	84.4	32.4	63.0	75.0
CAMP (Faster R-CNN, ResNet) [46]	50.1	82.1	89.7	39.0	68.9	80.2
SCAN (Faster R-CNN, ResNet) [21]	50.4	82.2	90.0	38.6	69.3	80.4
MMCA	**54.0**	**82.5**	**90.7**	**38.7**	**69.7**	**80.8**

Table 4. The impact of hidden space dimensionality on Flickr30K.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
64d	71.4	91.0	95.3	52.5	79.6	87.2
128d	72.6	91.8	96.5	53.5	80.4	86.8
256d	**74.2**	**92.8**	**96.4**	**54.8**	**81.4**	87.8
512d	73.6	91.7	95.5	53.9	81.2	**88.4**

Table 5. The impact of hidden space dimensionality on MS-COCO 1K testing set.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
64d	71.7	92.3	95.6	57.5	84.3	92.2
128d	73.7	93.8	97	60.7	**90.8**	94.7
256d	**74.8**	**95.6**	**97.7**	**61.6**	89.8	**95.2**
512d	74.1	94.2	97.5	60.4	88.5	94.3

Table 6. The impact of different values of α on Flick30K.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$\alpha = 0.1$	71.7	91.6	95.9	53.8	81.2	**89.1**
$\alpha = 0.2$	**74.2**	**92.8**	**96.4**	**54.8**	**81.4**	87.8
$\alpha = 0.4$	71.8	91.9	95.0	50.9	76.6	84.1
$\alpha = 0.8$	67.1	87.0	91.9	48.5	71.9	78.3

Furthermore, the qualitative results from image to text retrieval and the text to image retrieval on Flickr30K are illustrated in Figure 4 and Figure 5, respectively. For each image query, we show the top-5 retrieved sentences ranked by the similarity scores predicted by our approach. The correct retrieval sentences are marked in green, while the false ones are in black. It's clear that even the 'wrong' retrieval results have something similar to the queries, which proves our method does have extracted the interactions between visual and textual fragments. Figure 5 illustrates the qualitative results of image retrieval, with only one ground-truth image for each sentence. The top-4 retrieved images are sorted from left to right according to their similarity scores. We mark the true matches with green boxes. Once again, the top ranked retrieval results are reasonable.

5. Conclusions

In this paper, we propose a novel image and sentence matching method by jointly modeling both inter-modality and intra-modality relationships in a unified deep model. We first extract salient image regions and sentence tokens. Then we apply the proposed self-attention module and the cross-attention module to exploit the complex fine-grained relationships among the fragments. Finally, we update visual and textual features into a common embedding space by minimizing the hard-negative-based triplet loss. We have

Figure 4. Qualitative sentence search results using image query on Flickr30K. We show the top five retrieval sentences in descending relevant scores. The sentences in green are the true matches, and the descriptions in black are the false matches.

Figure 5. Qualitative image search results using sentence query on Flickr30K. We show the top four retrieval images in descending relevant scores. The images in green boxes are the true matches, and those in red boxes are the false matches.

systematically studied the influence of our idea and carried out experiments.The results demonstrate the effectiveness of our model by achieving significant performance.

6. Acknowledgment

References

[1] 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*, pages 6077–6086, 2018.

[2] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[6] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[8] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[9] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6639–6648, 2019.

[10] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, pages 7181–7189, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, pages 2310–2318, 2017.

[13] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, pages 6163–6171, 2018.

[14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[15] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[16] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[19] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, pages 4437–4446, 2015.

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[22] Guy Lev, Gil Sadeh, Benjamin Klein, and Lior Wolf. Rnn fisher vectors for action recognition and image annotation. In *ECCV*, pages 833–850. Springer, 2016.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[24] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[25] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014.

[26] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, pages 299–307, 2017.

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[28] Yuxin Peng and Jinwei Qi. Cm-gans: cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):22, 2019.

[29] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–264, 2018.

[30] Guoping Qiu. Indexing chromatic and achromatic patterns for content-based colour image retrieval. *Pattern Recognition*, 35(8):1675–1686, 2002.

[31] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[33] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[36] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[38] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[40] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

[41] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[42] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

[43] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[44] Shuhui Wang, Yangyu Chen, Junbao Zhuo, Qingming Huang, and Qi Tian. Joint global and co-attentive representation learning for image-sentence retrieval. In *ACM Multimedia*, pages 1398–1406, 2018.

[45] Yequan Wang, Minlie Huang, Li Zhao, et al. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615, 2016.

[46] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5764–5773, 2019.

[47] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[48] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2088–2096. ACM, 2019.

[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[50] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, pages 3441–3450, 2015.

[51] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[52] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

[53] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *arXiv preprint arXiv:1805.03508*, 2018.

[54] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–701, 2018.

[55] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535*, 2017.