Learning Hierarchical Semantic Correspondences for Cross-Modal Image-Text Retrieval

Sheng Zeng
Jiangxi Normal University
Nanchang, China
darrylzs@163.com

Changhong Liu*
Jiangxi Normal University
Nanchang, China
liuch@jxnu.edu.cn

Jun Zhou
Griffith University
Brisbane, Queensland, Australia
jun.zhou@griffith.edu.au

Yong Chen
Nanchang Institute of Technology
Nanchang, China
343225870@qq.com

Aiwen Jiang
Jiangxi Normal University
Nanchang, China
jiangaiwen@jxnu.edu.cn

Hanxi Li
Jiangxi Normal University
Nanchang, China
lihanxi2001@foxmail.com

ABSTRACT

Cross-modal image-text retrieval is a fundamental task in information retrieval. The key to this task is to address both heterogeneity and cross-modal semantic correlation between data of different modalities. Fine-grained matching methods can nicely model local semantic correlations between image and text but face two challenges. First, images may contain redundant information while text sentences often contain words without semantic meaning. Such redundancy interferes with the local matching between textual words and image regions. Furthermore, the retrieval shall consider not only low-level semantic correspondence between image regions and textual words but also a higher semantic correlation between different intra-modal relationships. We propose a multi-layer graph convolutional network with object-level, object-relational-level, and higher-level learning sub-networks. Our method learns hierarchical semantic correspondences by both local and global alignment. We further introduce a self-attention mechanism after the word embedding to weaken insignificant words in the sentence and a cross-attention mechanism to guide the learning of image features. Extensive experiments on Flickr30K and MS-COCO datasets demonstrate the effectiveness and superiority of our proposed method.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**.

KEYWORDS

cross-modal retrieval, image-text retrieval, fine-grained matching, self-attention, graph convolutional network

ACM Reference Format:
Sheng Zeng, Changhong Liu, Jun Zhou, Yong Chen, Aiwen Jiang, and Hanxi Li. 2022. Learning Hierarchical Semantic Correspondences for Cross-Modal Image-Text Retrieval. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR ’22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3512527.3531358

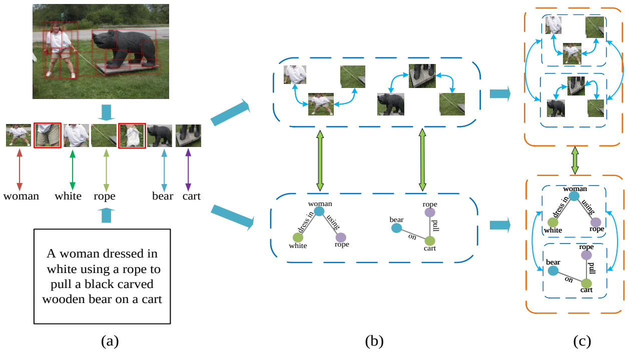

woman
white
rope
bear
cart
A woman dressed in white using a rope to pull a black carved wooden bear on a cart
woman
dress in
using
white
rope
bear
pull
on
cart
(a)
(b)
(c)

Figure 1: There are three different levels of image-text matching, where (a) indicates object-level matching, (b) indicates object relation-level matching, and (c) indicates higher-level semantic matching.

1 INTRODUCTION

With the development of the mobile Internet, different forms of media data such as texts, images, and videos are growing rapidly, making cross-modal image-text retrieval a fundamental research problem in information retrieval. Cross-modal retrieval [38] uses data in one modal as the query to retrieve data in another modal with the same or similar semantics. Although data in different modalities are heterogeneous and from multi-sources, they have common characteristics and are semantically interrelated when describing the same objects, concept and their relationships. Therefore, the key to cross-modal image text retrieval is to address both heterogeneity and the cross-modal semantic relevance between data of different modalities, which is still a very challenging research problem despite the significant progress in recent years.

Existing cross-modal image-text retrieval methods focus on exploring the semantic relevance of images and texts. Coarse-grained matching methods [16, 22, 44] map image and text directly into a

*Changhong Liu is the corresponding author.

ICMR ’22, June 27–30, 2022, Newark, NJ, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531358

common latent semantic space and then calculate the similarity between image and text in this common latent space. However, these methods only coarsely capture the global semantic correlation between different modal data and cannot effectively describe the local semantic correlation between image regions and text words. To address this drawback, fine-grained matching methods [3, 7, 18, 24, 25] model local similarity between image regions in the image and words in the text, and further fuse them to obtain global similarity metrics. Therefore, fine-grained matching methods have effectively improved the accuracy of cross-modal retrieval.

The challenge in cross-modal image-text retrieval also comes from the fact that a single sentence cannot fully describe all objects in the corresponding image in most cases. Image information redundancies often occur in the local matching stage, i.e. the corresponding words for some image regions cannot be found in the sentence, as shown in Figure 1(a). The sentence does not contain any words corresponding to image regions in red boxes. Moreover, the sentences often contain words that have no semantic meaning, such as "There", "are", and "a" in the sentence "There are two people, a man and a woman, sitting on a bus". Therefore, these redundancies interfere with the exact matching between image and text in the low-level matching stage. Additionally, cross-modal image-text matching considers not only low-level semantic correspondence between image regions and textual words but also the higher-level semantic correlation between different intra-modal relationships, as shown in Figure 1(b) and 1(c). Therefore, it is essential to understand both higher-level semantics and lower-level relationship between inter-modal objects, such as "woman" and "rope", "cart" and "bear" in Figure 1(b), and "woman-white-rope" and "bear-cart-rope" in Figure 1(c).

To address the above challenges and to achieve a finer and more accurate matching, we use the self-attention mechanism to strengthen words with semantic meaning and weaken words with no semantic meaning during the text feature extraction. In addition, we apply the cross-modal attention mechanism to filter out redundant image regions in the image. Further, we propose a multi-layer graph convolutional network to learn both the intra-modal relations of image regions and textual words and the relations of intra-modal relations. This enables object-level, object-relational level and higher semantic level matching, as illustrated in Figure 1(a), (b), and (c).

The major contributions of this work can be summarized as follows:

- We design a hierarchical semantic learning model (HSLM) with a multi-layer graph convolutional network (MGCN) to learn object-level, object-relational-level, and higher-level semantic features to achieve hierarchical semantic matching by both local and global alignment.
- To enhance image features, we propose a cross-modal attention module that combines a self-attention mechanism with a cross-modal attention mechanism to filter out insignificant words from the text and then enhance the image features using the text features.
- We verify our method on two standard benchmark datasets including Flickr30K and MS-COCO. Experimental results show that our method outperforms compared methods on all datasets. The experimental analysis also well demonstrates the superiority and reasonableness of our method.

2 RELATED WORK

Our proposed method focuses on image-text retrieval and explores the potential correspondences between vision and language. Current methods for image-text retrieval can be broadly classified into two categories: (1) coarse-grained matching methods, which map the entire image and text into a common embedding space to learn global semantic correlations between modalities, and (2) fine-grained matching methods, which focus on learning local semantic correlations between image and text objects.

Coarse-grained matching method. Andrea et al. [11] made the first attempt to unify image features and text features by a linear mapping. Convolutional Neural Network (CNN) [23] and Long Short-term Memory Recurrent Network (LSTM) [15] are commonly used to extract the image and text features individually, and then embed them into a common semantic space to achieve cross-modal matching [14, 21]. The correlation between image and text can also be learned via autoencoder decoder [10] or recurrent residual fusion network [29]. Vendrov et al. [35], on the other hand, learn order-embeddings that is a mapping with order-preserving between the visual semantic hierarchy. Wang et al. [37] maximized the correlation of data from different modalities through linear projection. Faghri et al. [9] improved training strategy by hard negative mining to achieve a significant improvement. Although these methods achieved good results on the image-text retrieval task, they ignored the semantic association between image and text data at a fine-grained level.

Fine-grained matching method. Recently, many efforts have been devoted to exploring fine-grained correspondence for cross-modal image-text retrieval. To extract the local features of image regions and textual words, bottom-up attention model [1] based on Faster-RCNN [32] and bidirectional GRU(B-GRU) [5] are often used for image and text [3, 7, 18, 24, 25]. Nevertheless, these approaches don't consider the intra-modal relations of image regions and textual words. To solve this problem, Graph Convolutional Network (GCN) [20] is introduced to learn the intra-modal relations between image regions, and then fused the extracted local features as the global representation in VSRN [25] and DSRAN [42]. Additionally, the methods based on scene-graph [8, 28] learn the intra-modal relations and fine-grained correspondences by constructing a visual scene graph and a textual scene graph. To better capture intra-modal relations and inter-modal semantic correspondence, the hierarchical matching methods that simultaneously learn local and global correpondences are growing. With the pre-trained BERT as the textual encoder, DIME [31] utilizes the routing mechanism to form a novel hierarchical network for intra-modal reasoning and inter-modal alignment. SGRAF [7] and SHAN [17] learn the correspondences between local and global cross-modal pairs by local and global alignments, which are somewhat similar to our work. But we consider the importance of textual words and image regions, explicitly construct the object-level, object-relational-level and higher-level learning networks to form a multi-layer graph convolutional network, and then learn hierarchical semantic correspondences by local alignment and global alignment.

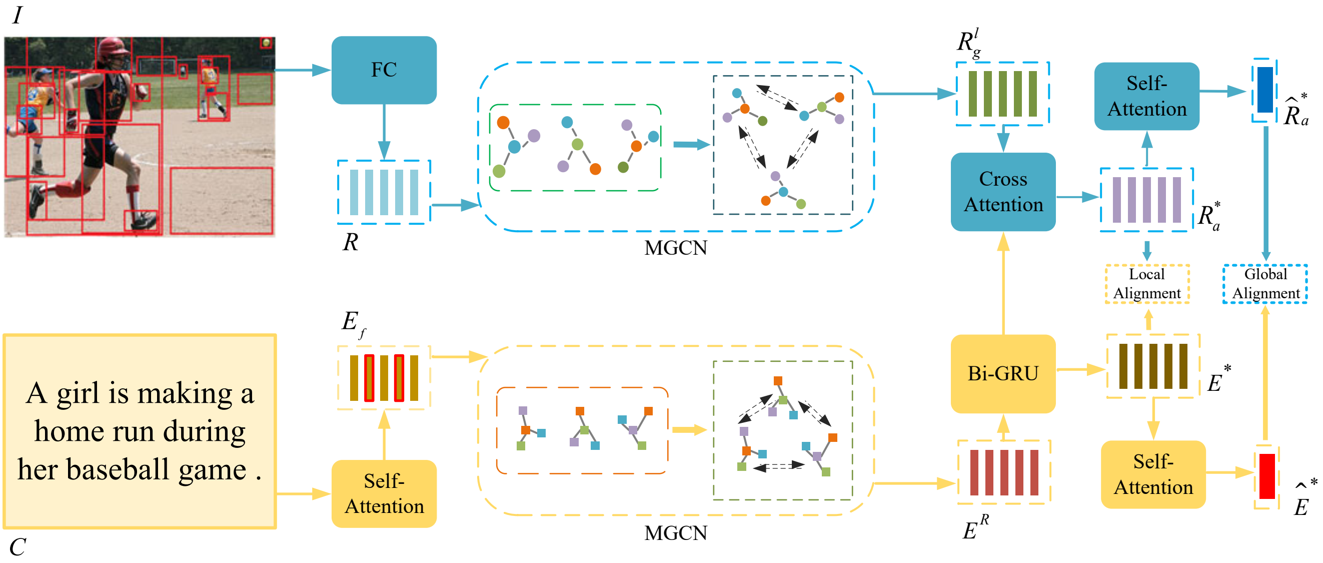


Figure 2: The framework of the proposed method.

Attention Mechanism. The attention mechanism is effective in filtering and enhancing data features, which has been successfully adopted in natural language processing [6, 30, 33, 34] and image processing [12, 43, 45, 48, 49]. It has also been introduced in recent cross-modal image-text retrieval tasks. Different attention mechanisms are used in intra-modal interactions, inter-modal interactions, or jointly. Self-attention are usually utilized to capture intra-modal semantic correlations [2, 36, 39], whereas cross-attention are exploited to learn local inter-modal latent alignment [24, 27, 40]. Yu et al. [47] also apply a stack of self-attention blocks into learning local inter-modal correlations. Ji et al. [17] implement local and global alignments by multiple cross-attentions. Self-attention and cross-attention play an important role in image-text matching. They have also been combined into learning intra-modal relations within each modality and inter-modal semantic alignments synchronously [4, 41, 50]. Similarly, we integrate cross-attention and self-attention mechanisms into our proposed model. Different from most methods that learn intra-modal correlations based on the self-attention mechanism after GRU, we use the self-attention mechanism to filter out words that have no semantic meaning. Then we apply the cross-attention mechanism to guide the learning of image features. This strategy helps to retain informative and meaningful words.

3 PROPOSED METHOD

In this section, we describe the details of the hierarchical semantic learning model (HSLM) for cross-modal image-text retrieval. Figure 2 shows the framework of our proposed method. Our goal is to learn hierarchical semantic correspondences between image and text. The self-attention mechanism is firstly used to enhance textual word features. For texts, based on these extracted representations of the textual words, we model multi-level intra-modal relationships between textual words using MGCN. The learned local features are fed into GRU for contextual relationships building and then fused into global features using the attention mechanism. For images, we obtain image regions and their features generated by the Faster-RCNN. We also use MGCN to learn multi-level intra-modal relationships between image regions. After that, the local text features with the relationships guide the learning of the local image features by the cross-attention mechanism. Then we fuse the learned local semantic features to obtain the global features of the image. Finally, the local and global features of the image and the text are matched and further combined to calculate the image-text similarity. The details are introduced as follows.

3.1 Image Representation Extraction

Given an image I, we use a bottom-up attention model [1] based on Faster-RCNN [32] to select the top N image region features with the highest class detection confidence score. A fully connected layer is then applied to transform them to a d_1-dimensional embedding as local object-level image representations, denoted as:

$$R = \left\{ r_i | i = 1, ..., N, r_i \in \mathbb{R}^{d_1} \right\}, \tag{1}$$

where r_i represents the i^{th} image region feature, $N = 36$ and d_1 is the dimension of a single image region feature.

3.2 Text Representation Extraction

For a text C containing M words, each word w_j is represented by a continuous embedding vector $e_j = W_e w_j, \forall j \in [1, M]$, where W_e is the embedding matrix to be learned and d_2 denotes the dimension of the word feature. Thus the word object-level features of text sentence can be described as:

$$E = \left\{ e_j | j = 1, ..., M, e_j \in \mathbb{R}^{d_2} \right\}. \tag{2}$$

Note that sentences often contain some words that have no semantic meaning. In order to strengthen the words with real meaning and weaken the words without real meaning, we use a self-attention mechanism to strengthen and filter the textual word features. Specifically, we firstly compute the key E_K, query E_Q, and value E_V in the self-attention mechanism:

$$E_K = W_K E, \tag{3}$$

$$E_Q = W_Q E, \tag{4}$$

$$E_V = W_V E, \tag{5}$$

where W_K, W_Q, and W_V are the parameter matrices that need to be learned by back propagation. Next, the enhanced word object-level features of the text can be defined as:

$$E_f = (\frac{E_K E_Q}{M}) E_V. \tag{6}$$

Then the word object-level features of the text are denoted as $E_f = \{e_j^f | j = 1, ..., M, e_j^f \in \mathbb{R}^{d_2}\}$, where e_j^f is the feature representation of the j^{th} word and M is the total number of words in a sentence.

3.3 Higher-level Representation Learning

Both text and image have direct or indirect semantic relationships between intra-modal objects, including object-to-object relations and relation-to-relation relationships (higher-level semantic relationships). In order to capture these intra-model relationships, we construct two fully connected graphs for image regions and textual words separately and learn multi-level relationship representations by multi-layer graph convolutional network (MGCN).

For Image. Given the local object-level image features R defined in Equation 1, following [25], we construct a relation graph $G = (R, A)$ for the intra-modal objects of the image, where A is the relation matrix between image objects. The relation A_{ij} between the i^{th} object and the j^{th} object is defined as:

$$A_{ij} = \mu(r_i)^T \nu(r_j), \tag{7}$$

where $\mu(.)$ and $\nu(.)$ are two fully connected layers learned by back propagation.

Without loss of generality, GCN is used to learn the relationships between the graph objects to capture the intra-modal object relationship information in the image. The learned image object-relational-level features with intra-modal object relationship information are denoted as:

$$R_g = W_r(ARW_g) + R, \tag{8}$$

where W_g and W_r are the weight matrices of the GCN.

As we need to learn higher-level semantics, we propose a multi-layer graph convolutional network (MGCN) that further feeds the image object-relational-level R_g into the MGCN to learn higher-level semantic relationships, as follows:

$$R_g^l = W_h^l(A^l R_g^{l-1} W_m^l) + R_g^{l-1}, \tag{9}$$

where l is the number of layers of MGCN, W_h^l and W_m^l denote the weight matrix in the l^{th} layer, A^l and R_g^l are the relationship matrix and features in the l^{th} layer, respectively.

Using Equations 7-9, we can obtain local image features with multiple semantic information (object-level, object-relational-level, higher-level):

$$R^* = \left\{ r_i^* | i = 1, ..., N, r_i^* \in \mathbb{R}^{d_1} \right\}. \tag{10}$$

For Text. In the same way, as we learn the relationship between image region features, the word object-level features E_f obtained from Equation 6 are fed into the MGCN to learn the higher-level semantic features in the text. Thus, local text features with multi-layer semantic information can be represented as:

$$E^R = \{e_j^r | j = 1, ..., M, e_j^r \in \mathbb{R}^{d_2}\}. \tag{11}$$

Considering that there is a strong sequential relationship between words in the text, we use bidirectional GRU (B-GRU) [5] to enhance local text features with contextual relationships:

$$\begin{matrix} \overrightarrow{h}_j = \overrightarrow{GRU}(e_j^r, \overrightarrow{h}_{j-1}) \\ \overleftarrow{h}_j = \overleftarrow{GRU}(e_j^r, \overleftarrow{h}_{j+1}) \end{matrix}, \tag{12}$$

where $\overrightarrow{h}_j$ and $\overleftarrow{h}_j$ are the j^{th} hidden states of forward $\overrightarrow{GRU}$ and backward $\overleftarrow{GRU}$, respectively. Thus, the feature vector of the j^{th} word can be described as:

$$t_j = \frac{\overrightarrow{h}_j + \overleftarrow{h}_j}{2}. \tag{13}$$

Local text features with multi-layer semantic information can be denoted as:

$$E^* = \left\{ t_j | j = 1, ..., M, t_j \in \mathbb{R}^{d_1} \right\}. \tag{14}$$

3.4 Cross Attention Module

In most cases, a sentence cannot fully describe the content or objects in the corresponding image. For example, in the benchmark datasets Flickr30K and MS-COCO adopted in this research, each image corresponds to 5 sentences. Therefore at the local matching step, the image features may contain some redundant information. To highlight some informative image regions, we design the cross-attention module guided by the importance and relations of the textual words.

We first provide a general formulation of an attention mechanism designed for the cross-attention problem. Given local image features R^* and E^* obtained from Equations 10 and 14 respectively, we define the cross attention process as:

$$R_a^* = f(E^*, R^*) \odot R^*, \tag{15}$$

$$f(x, y) = sigmoid(x \odot y), \tag{16}$$

where $f(x, y)$ is the attention function to calculate the scores for each pair t_j and r_i^*. x and y represent the data from two different modes, and R_a^* is the local enhanced image features. $\odot$ indicates the dot product operation.

3.5 Similarity Calculation

So far we have obtained the local features E^* of the text and the local features R_a^* of the image, respectively. Subsequently, we average the local features of the text, denoted as $\hat{E}$. We apply a self-attention mechanism [34], using $\hat{E}$ as a query, to enhance the local features and further fuse them to obtain the global representation $\hat{E}^*$. Similarly, the global image representation $\hat{R}_a^*$ is computed by the self-attention method over all of image region features R_a^*.

we follow [7] to compute the similarity representation whose function is defined as:

$$S(m_1, m_2) = \frac{W|m_1 - m_2|^2}{||W|m_1 - m_2|^2||_2}, \tag{17}$$

where m_1, m_2 are two vectors from different modalities. $|.|^2$and $||.||_2$ indicate element-wise square and l_2 norm respectively, and W is a learnable parameter matrix.

The local and global similarity representations between image and text are defined as S_L and S_G:

$$S_L = S(E^*, R_a^*), \tag{18}$$

$$S_G = S(\hat{E}^*, \hat{R}_a^*). \tag{19}$$

Afterwards, we concatenate the local and global similarity representations and then enhance the importance through weighting:

$$S^* = S_L || S_G, \tag{20}$$

$$S_A = Sigmoid(W_s S^*) \odot S^*, \tag{21}$$

where || indicates concatenation, S^* denotes the similarity representation with multiple-layer semantic information after concatenation, W_s is the weight matrix learned by back propagation, $\odot$ indicates the dot product operation, and S_A is the enhanced similarity representation.

Finally, the similarity representation is transformed into a similarity score between image and text through a fully connected layer.

$$S_F = FC(S_A), \tag{22}$$

where FC indicates a fully connected layer and S_F is the similarity score between image and text.

3.6 Alignment Objective

We utilize the bi-directional triplet ranking loss [9] which emphasizes hard negatives [24], i.e., the negatives closest to each training query. The loss function can be defined as:

$$\begin{aligned} L =& [\beta - S(I, C) + S(I, \widehat{C})]_+ + \\ & [\beta - S(I, C) + S(\widehat{I}, C)]_+, \end{aligned} \tag{23}$$

where β is the margin parameter, the operator $[z]_+ = max(0, z)$ compares the tolerance value with zero, (I, C) is a positive image-text pair, $\widehat{C} = \arg\max_{d \neq C} S(I, d)$ and $\widehat{I} = \arg\max_{k \neq I} S(k, C)$ stand for the hardest negatives for a positive pair (I, C). During the model training, instead of summing over the hardest negatives in the entire training set, we only use the hard negatives in each mini-batch.

4 EXPERIMENTS

In order to demonstrate the effectiveness as well as the scalability of the proposed method, we conduct experiments on two commonly used publicly available datasets, Flickr30K [46] and MS-COCO [26], for Image-to-Text and Text-to-Image retrieval tasks. Following the related works, we take Recall@K (R@K, K = 1, 5, 10) as the evaluation metric, i.e. the fraction of queries that retrieve the correct items in the top-K results. We also use the Rsum metric defined by the sum of recall metrics at K = 1, 5, 10, to measure the overall performance of retrieval.

Table 1: Quantitative evaluation results of two retrieval tasks on Flickr30K test set in terms of Recall@K (R@K)

	Image-to-Text			**Text-to-Image**			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
SCAN [24]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
CAMP [40]	68.1	89.7	95.2	51.5	77.1	85.3	466.9
BFAN [27]	68.1	91.4	-	50.8	78.4	-	288.7
PFAN [39]	70.0	91.8	95.0	50.4	78.7	86.1	472.0
CVSE [36]	73.5	92.1	95.8	52.9	80.4	87.8	482.5
VSRN [25]	71.3	90.6	96.0	54.7	81.8	88.2	482.6
DP-RNN [4]	70.2	91.6	95.8	55.5	81.3	88.2	482.6
IMRAM [3]	74.1	93.0	96.6	53.9	79.4	87.2	484.2
MMCA [41]	74.2	92.8	96.4	54.8	81.4	87.8	487.4
GSMN [28]	76.4	94.3	97.3	57.4	82.3	89.0	496.8
CAAN [50]	70.1	91.6	97.2	52.8	79.0	87.9	478.6
DSRAN(GRU) [42]	79.6	95.6	**97.5**	58.6	**85.8**	**91.3**	508.4
HAN[47]	74.1	92.4	96.4	54.8	81.1	87.4	486.2
SHAN [17]	74.6	93.5	96.9	55.3	81.3	88.4	490.0
HSGMP [8]	73.4	93.0	96.8	55.0	81.4	88.2	487.8
SGRAF [7]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
ours(HSLM)	**79.9**	**95.7**	**97.5**	**60.7**	84.7	90.1	**508.6**
GRAN [2]	76.6	93.6	97.6	60.5	86.1	91.3	505.7
DIME [31]	$\underline{81.0}$	$\underline{95.9}$	$\underline{98.4}$	$\underline{63.6}$	$\underline{88.1}$	$\underline{93.0}$	$\underline{520}$

4.1 Datasets and Protocols

Flickr30K [46] consists of 31,783 images. Each image is paired with five text descriptions. We split the dataset into 1,000 images for validation, 1,000 images for testing, and the rest for training as in [18]. **MS-COCO** [26] contains 123,287 images, and each image is annotated with five textual sentences. Following the split protocol of [13], 113,287 images are used as the training set, 5,000 images are used as the validation set, and 5,000 images are taken as the test set. In order to verify the stability and robustness of the model, the experimental results are reported in two testing ways. The first way (MS-COCO 5K) directly uses 5,000 test images for testing. The second way (MS-COCO 1K) uses a 5-fold validation approach for testing, where 1,000 images are tested each time, and then the average of 5 test results is taken as the final test result.

4.2 Implementation Details

In the experiments, we use the Faster-RCNN model to extract the top 36 image regions in terms of confidence score. The dimension of the joint embedding space d_1 is 1024 and the dimension of word features d_2 is 300. After experiments, it was found that the best experimental results were obtained when the number of layers l of MGCN was set to 4. Adam [19] is employed as the optimizer during model training. For the Flickr30K dataset, the initial learning rate is set to 0.0002 and the learning rate decay with a factor of 0.1 is performed every 20 iterations. For the MS-COCO dataset, the initial learning rate is set to 0.0002 and the learning rate decay with a factor of 0.1 is performed every 15 iterations. The margin β is empirically set as 0.2 in Equation 23. The number of iterations (epoch) is set to 40 and the batch size is 128 for both Flickr30K and MS-COCO. For the evaluation on the test set, we save the best performing model on the verification set every time. The best model

Table 2: Quantitative evaluation results of two retrieval tasks on MS-COCO 1K test set in terms of Recall@K (R@K)

	Image-to-Text			Text-to-Image			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
SCAN [24]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
CAMP [40]	72.3	94.8	98.3	58.5	89.7	95.0	506.8
BFAN [27]	74.9	95.2	-	59.4	88.4	-	317.9
PFAN [39]	76.5	96.3	99.0	61.6	89.6	95.2	518.2
CVSE [36]	74.8	95.1	98.3	59.9	89.4	95.2	512.7
VSRN [25]	76.2	94.8	98.2	62.8	89.7	95.1	516.8
DP-RNN [4]	75.3	95.8	98.6	62.5	89.7	95.1	517
IMRAM [3]	76.7	95.6	98.5	61.7	89.1	95.0	516.6
MMCA [41]	74.8	95.6	97.7	61.6	89.8	95.2	514.7
CAAN [50]	75.5	95.4	98.5	61.3	89.7	95.2	515.6
GSMN [28]	78.4	96.4	98.6	63.3	90.1	95.7	522.5
DSRAN(GRU) [42]	**80.4**	**96.7**	98.7	64.2	90.4	95.8	526.2
HAN[47]	78.7	96.4	**98.8**	**65.4**	90.5	95.3	525.1
SHAN [17]	76.8	96.3	98.7	62.6	89.6	95.8	519.8
HSGMP [8]	76.4	95.4	98.3	63.0	88.8	94.7	516.6
SGRAF [7]	79.6	96.2	98.5	63.2	90.7	**96.1**	524.3
ours(HSLM)	80.3	96.5	98.7	65.0	**90.9**	96.0	**527.3**
GRAN [2]	79.1	96.6	98.8	65.2	91.4	96.3	527.4
DIME [31]	78.8	96.3	98.7	64.8	91.5	96.5	526.6

is determined based on the sum of the recalls in the validation set. All experiments in this work are implemented on one RTX2080TI GPU and the Pytorch framework.

4.3 Comparison with the State-of-the-art Methods

We compare our proposed method with the following state-of-the-art methods: (1) methods that focus on local region-word correspondence, including SCAN [24], PFAN [39], and IMRAM [3]; (2) methods with the intra-modal relation learning, including VSRN [25] and DSRAN [42]; (3) methods that learn local and global correpondences, including GSMN [28] and HSGMP [8], SGRAF [7], SHAN [17], DIME [31]; (4) methods based on the attention mechanism, including CAMP [40], BFAN [27], CVSE [36], HAN[47], GRAN [2], MMCA [41], CAAN [50] and DP-RNN [4].

Most of these approaches utilized Faster-RCNN as image backbone with 36 regions, and GRU as text backbone. But GRAN [2] and DIME [31] used an additional pre-trained BERT [6] as the text encoder, and DSRAN [42] added the global image features encoded by a ResNet152.

Results on Flickr30K Table 1 shows the quantitative results of different methods on the Flickr30K dataset. Our method achieved 79.9%, 95.7%, 97.5% on Image-to-Text retrieval task and 60.7%, 84.7%, and 90.1% on Text-to-Image retrieval for R@1, R@5, and R@10, respectively. Our method outperforms other state-of-the-art methods except for DIME [31] with the pre-trained BERT model and DSRAN with the global feature encoded by a ResNet152 at all of metrics on two retrieval tasks. Moreover, compared with the recent GRAN [2] with the pre-trained model, the performance gains at R@1 and R@5 are 3.3 and 2.1 respectively for Image-to-Text retrieval. This also reveals the importance of local correspondence.

Table 3: Quantitative evaluation results of two retrieval tasks on MS-COCO 5K test set in terms of Recall@K (R@K)

	Image-to-Text			Text-to-Image			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
SCAN [24]	50.4	82.2	90.0	38.6	69.3	80.4	410.9
CAMP [40]	50.1	82.1	89.7	39.0	68.9	80.2	410
VSRN [25]	53.0	81.1	89.4	40.5	70.6	81.1	415.7
IMRAM [3]	53.7	83.2	91.0	39.7	69.1	79.8	416.5
MMCA [41]	50.4	82.5	90.7	38.7	69.7	80.8	416.7
CAAN [50]	52.5	83.3	90.9	41.2	70.3	**82.9**	421.1
DSRAN(GRU) [42]	57.6	85.6	91.9	41.5	71.9	82.1	430.6
HSGMP [8]	53.3	83.7	90.3	41.5	69.6	80.3	418.7
SGRAF [7]	57.8	-	91.6	41.9	-	81.3	272.6
ours(HSLM)	**59.9**	**85.8**	**92.7**	**43.3**	**72.1**	82.6	**436.3**
GRAN [2]	58.2	85.0	91.7	43.2	73.0	83.1	434.2
DIME [31]	59.3	85.4	91.9	43.1	73.0	83.1	435.8

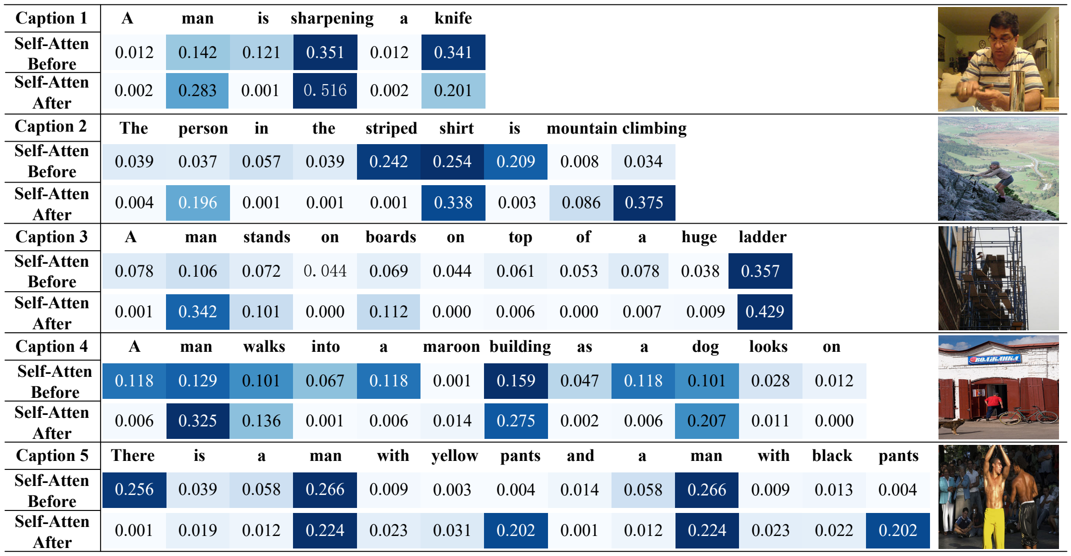
The results show the methods with intra-modal relation learning (e.g., VSRN [25] and DSRAN [42]) produce better performances than the methods that focus on local region-word correspondence (e.g., SCAN [24], PFAN [39], and IMRAM [3]). Compared with VSRN, our proposed method learns both local and global correspondences to make further improvement, as did by GSMN, HSGMP, SGRAF, SHAN, and DIME. Similar to CAAN [50], MMCA [41] and SGRAF [7], the proposed method captures the intra-modal relations and inter-modal semantic alignments using self-attention and cross-attention, but the self-attention is utilized to filter out the insignificant words and the cross-modal attention simultaneously strengthens local image features in our method. The operations highlight the local correspondence, so that our proposed method generates better results and improves 2.6%, 1.7%, and 0.1% on the image-to-Text retrieval and 3.7%, 2%, and 1.4% on the Text-to-Image retrieval at R@1, R@5, and R@10 over SGRAF.

Results on MS-COCO. The experimental quantitative results on the MSCOCO 1K test set are shown in Table 2. Our method outperforms most of recent methods and achieves the highest performance in Rsum. In addition, the experimental validation is also carried out on the MS-COCO 5K test set and the results are shown in Table 3. From the results, we can see that our method is remarkably superior to the state-of-the-art methods in most metrics. When measured by R@1 and R@10, our method outperforms the current state-of-the-art method with 3.6% and 1.2% on the Image-to-Text retrieval, and 3.3% and 1.5% on the Text-to-Image retrieval relatively. Moreover, our method surpasses DIME [31] and GRAN [2] with the pre-trained BERT model. This further shows the effectiveness of our method and indicates its ability to learn hierarchical semantic correlation between the image and the text through MGCN and cross-attention.

From the above experimental results, we can see that our proposed method not only performs well on the small dataset Flickr30K but also outperforms most related methods on the large dataset MS-COCO, which fully demonstrates the superiority and scalability of our proposed method.

Table 4: Ablation studies on the Flickr30K test set. Results are reported in terms of Recall@K (R@K). “Self-Attn” refers to the model with self-attention module, “Cross-Attn” represents the model with cross-attention module and "MGCN" means the model with MGCN module. "Glob" and "Loc" refer to using global alignment or local alignment.

	Model Settings					Image-to-Text			Text-to-Image			
Model	MGCN	Self-Attn	Cross-Attn	Glob	Loc	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
1				✓		58.7	85.3	91.5	43.5	72.0	80.0	431.0
2		✓		✓		60.1	85.6	91.2	45.0	74.6	82.7	437.5
3	✓			✓		61.4	84.3	91.9	45.5	73.8	81.8	438.7
4	✓	✓		✓		60.9	86.8	93.6	45.8	74.7	82.5	444.2
5			✓		✓	72.1	91.8	96.1	54.2	80.3	86.5	481.0
6		✓	✓		✓	71.8	92.6	96.3	54.9	80.8	86.8	483.2
7	✓		✓		✓	73.8	92.1	96.1	55.8	80.5	86.7	485.0
8	✓	✓	✓		✓	72.7	92.8	97.3	56.5	81.9	88.3	489.5
9			✓	✓	✓	72.7	92.1	96.0	55.3	81.0	87.5	484.6
10		✓	✓	✓	✓	72.2	93.4	96.8	55.8	81.6	87.5	487.3
11	✓		✓	✓	✓	75.2	94.1	96.9	57.1	**82.9**	**88.5**	494.6
12	✓	✓	✓	✓	✓	**77.1**	**94.1**	**97.4**	**58.1**	82.6	88.4	**497.8**



Caption 1	**A**	**man**	**is**	**sharpening**	**a**	**knife**							
Self-Atten Before	0.012	0.142	0.121	0.351	0.012	0.341							
Self-Atten After	0.002	0.283	0.001	0. 516	0.002	0.201							
Caption 2	**The**	**person**	**in**	**the**	**striped**	**shirt**	**is**	**mountain**	**climbing**				
Self-Atten Before	0.039	0.037	0.057	0.039	0.242	0.254	0.209	0.008	0.034				
Self-Atten After	0.004	0.196	0.001	0.001	0.001	0.338	0.003	0.086	0.375				
Caption 3	**A**	**man**	**stands**	**on**	**boards**	**on**	**top**	**of**	**a**	**huge**	**ladder**		
Self-Atten Before	0.078	0.106	0.072	0. 044	0.069	0.044	0.061	0.053	0.078	0.038	0.357		
Self-Atten After	0.001	0.342	0.101	0.000	0.112	0.000	0.006	0.000	0.007	0.009	0.429		
Caption 4	**A**	**man**	**walks**	**into**	**a**	**maroon**	**building**	**as**	**a**	**dog**	**looks**	**on**	
Self-Atten Before	0.118	0.129	0.101	0.067	0.118	0.001	0.159	0.047	0.118	0.101	0.028	0.012	
Self-Atten After	0.006	0.325	0.136	0.001	0.006	0.014	0.275	0.002	0.006	0.207	0.011	0.000	
Caption 5	**There**	**is**	**a**	**man**	**with**	**yellow**	**pants**	**and**	**a**	**man**	**with**	**black**	**pants**
Self-Atten Before	0.256	0.039	0.058	0.266	0.009	0.003	0.004	0.014	0.058	0.266	0.009	0.013	0.004
Self-Atten After	0.001	0.019	0.012	0.224	0.023	0.031	0.202	0.001	0.012	0.224	0.023	0.022	0.202

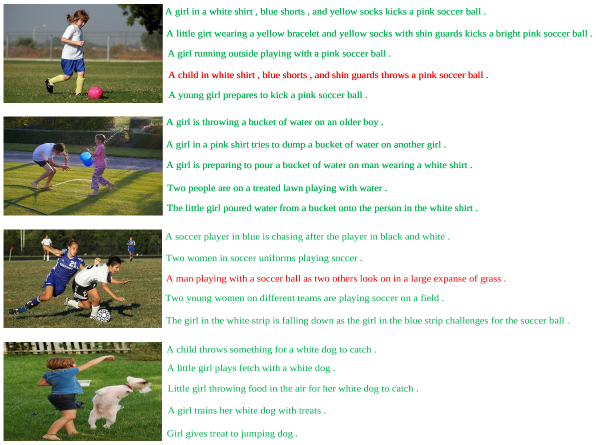
Figure 3: visualization of the word correlation in the sentence before self-attention and after self-attention using Flickr30K dataset. Self-Atten Before and Selft-Atten After separately refer to the word correlation values before self-attention and after self-attention. The left column is the title of each row, the middle is the word correlation values and the right column is the matching image for each sentence.

4.4 Ablation Studies

In order to validate the impact of different network structures, we progressively add different modules, including MGCN module, self-attention module, and cross-attention module, to the baseline model. We conduct the ablation experiments on the Flickr30K dataset with or without these modules. The results are shown in Table 4.

The model with self-attention module is superior to single global alignment baseline with about 2.0 R@1 gain and 6.5 Rsum gain in the first two lines. This shows that the self-attention mechanism is effective to focus on the significant words of the text. MGCN further improves the performance with about 7.7 Rsum gain by learning the intra-modal relations in lines 1 and 3. Nevertheless, combining MGCN with local alignment and cross-attention, the performance gain in Rsum is 47 in lines 3 and 7. This demonstrates that local alignment can be effectively integrated with MGCN features to learn multi-layer local semantic correspondences. The final line shows that the performance of the model has improved remarkably, because the self-attention and cross-attention modules can effectively learn the key features in text and image, and MGCN can acquire the higher-level semantic features of images. As a result, the retrieval accuracy of the model has increased significantly.

Furthermore, we analyze the effectiveness of self-attention on weakening insignificant words in the text. To illustrate this capability, we visualize the word correlation with other words before



A girl in a white shirt , blue shorts , and yellow socks kicks a pink soccer ball .
A little girt wearing a yellow bracelet and yellow socks with shin guards kicks a bright pink soccer ball .
A girl running outside playing with a pink soccer ball .
A child in white shirt , blue shorts , and shin guards throws a pink soccer ball .
A young girl prepares to kick a pink soccer ball .

A girl is throwing a bucket of water on an older boy .
A girl in a pink shirt tries to dump a bucket of water on another girl .
A girl is preparing to pour a bucket of water on man wearing a white shirt .
Two people are on a treated lawn playing with water .
The little girl poured water from a bucket onto the person in the white shirt .

A soccer player in blue is chasing after the player in black and white .
Two women in soccer uniforms playing soccer .
A man playing with a soccer ball as two others look on in a large expanse of grass .
Two young women on different teams are playing soccer on a field .
The girl in the white strip is falling down as the girl in the blue strip challenges for the soccer ball .

A child throws something for a white dog to catch .
A little girl plays fetch with a white dog .
Little girl throwing food in the air for her white dog to catch .
A girl trains her white dog with treats .
Girl gives treat to jumping dog .

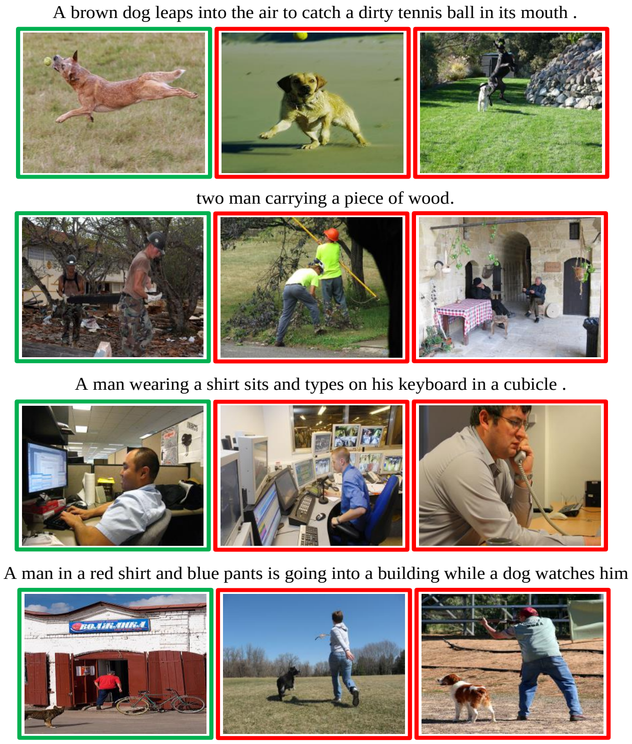
Figure 4: Visualization of the Image-to-Text retrieval on Flickr30K. For each image query, we display the first five texts, in which the mismatched texts are in red and the matched texts are in green.

self-attention and after self-attention using the Flickr30K dataset. The similarity matrix between textual words is firstly computed on the word embedding vector of the text, then the sum along each column is taken as the word correlation. All values in a sentence are normalized. The smaller the value, the less relevant it is to other words, which indicates less significance.

We list 5 sentences and mark the word correlation values using different levels of blue colors, as shown in Figure 3. From the middle column, we can observe that the word correlation values are more evenly distributed before self-attention, but after self-attention the bigger values mainly focus on the entity objects and action verbs, and the correlation values of some of unimportant words are decreased. In caption 1, for example, the correlation value of the word "is" is close to that of the word "man" before self-attention, yet the correlation values of the words "is" and "a" are greatly reduced and the correlation values of the word "man" and "sharpen" are underlined after self-attention. This indicates that the self-attention mechanism can successfully pay attention to more significant words.

4.5 Visualization and Analysis

Figure 4 shows the results of our method for image-to-text retrieval where the text in green font indicates the correct retrieval results and the text in red font indicates the incorrect retrieval results. Analysing the incorrect results for the first image retrieved in the figure shows that although the final retrieval result in line 4 is incorrect, the majority of the word objects in the text are correctly matched to the image regions in the image, such as "child", "in white shirt", "blue shorts", "pink soccer ball" and so on. This indicates that our method can match images and text at a fine-grained level.



A brown dog leaps into the air to catch a dirty tennis ball in its mouth .

two man carrying a piece of wood.

A man wearing a shirt sits and types on his keyboard in a cubicle .

A man in a red shirt and blue pants is going into a building while a dog watches him .

Figure 5: Visualization of the Text-to-Image retrieval on Flickr30K. For each text query, we display the top three ranked images from left to right, in which the mismatched images are in red boxes and the matched images are in green boxes.

In addition, we also visualize the results of text-to-image retrieval, where the green boxes indicate the correct retrieval results and the red boxes indicate the incorrect retrieval results. The retrieval results are shown in Figure 5. The retrieved images are relatively similar in terms of content and scenario, but our method is still able to retrieve the correct images accurately, which demonstrates the effectiveness of our proposed method.

5 CONCLUSION

In this paper, we have introduced a novel hierarchical semantic correspondence approach that models object-level correspondence, object-relational-level correspondence and higher-level semantic correspondence between image and text respectively. In addition, we have presented a cross-modal attention module that combines the self-attention mechanism with the cross-attention mechanism to guide the learning of image features using the hierarchical semantic relations of words in the text. Extensive experiments on two-widely used benchmark datasets, Flickr30K and MSCOCO, show that our proposed method outperforms the state-of-the-art methods on most of image-text retrieval metrics. In the future, we will consider adding a generation module that generates image from text or generates text from image to enhance the consistency of cross-modal retrieval.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under grants 62067004, 61966018 and 61962027.

REFERENCES

[1] 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.* 6077–6086.

[2] Jie Cao, Shengsheng Qian, Huaiwen Zhang, Quan Fang, and Changsheng Xu. 2021. Global relation-aware attention network for image-text retrieval. In *Proceedings of the International Conference on Multimedia Retrieval.* 19–28.

[3] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 12652–12660.

[4] Tianlang Chen and Jiebo Luo. 2020. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10583–10590.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[6] 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.* 4171–4186.

[7] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence.* 1218–1226.

[8] Yu Duan, Yun Xiong, Yao Zhang, Yuwei Fu, and Yangyong Zhu. 2021. HSGMP: Heterogeneous scene graph message passing for cross-modal retrieval. In *Proceedings of the 2021 International Conference on Multimedia Retrieval.* 82–91.

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference.* 12.

[10] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia.* 7–16.

[11] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems.* 2121–2129.

[12] Jianlong Fu, Heliang Zheng, and Tao Mei. 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4476–4484.

[13] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7181–7189.

[14] Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, and Chunhong Pan. 2016. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia* 18, 7 (2016), 1363–1377.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[16] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 6163–6171.

[17] Zhong Ji, Kexin Chen, and Haoran Wang. 2021. Step-wise hierarchical alignment network for image-text matching. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence.* 765–771.

[18] Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2017), 664–676.

[19] Diederik P. Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations.* 1–15.

[20] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations.* 1–14.

[21] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. In *Proceedings of the 2014 Conference on Neural Information Processing Systems workshop on Deep Learning.* 1–13.

[22] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4437–4446.

[23] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1, 4 (1989), 541–551.

[24] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision.* 212–228.

[25] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision.* 4654–4662.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision.* 740–755.

[27] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia.* 3–11.

[28] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 10921–10930.

[29] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision.* 4107–4116.

[30] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* 1412–1421.

[31] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1104–1113.

[32] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149.

[33] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* 379–389.

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

[35] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *Proceedings of the 4th International Conference on Learning Representations.* 1–12.

[36] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020. Consensus-aware visual-semantic embedding for image-text matching. In *Proceedings of the European Conference on Computer Vision (Lecture Notes in Computer Science, Vol. 12369).* 18–34.

[37] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 5005–5013.

[38] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision.* 1508–1517.

[39] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position focused attention network for image-text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.* 3792–3798.

[40] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. CAMP: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision.* 5763–5772.

[41] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 10941–10950.

[42] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. 2021. Learning dual semantic relations with graph attention for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 7 (2021), 2866–2879.

[43] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (Lecture Notes in Computer Science, Vol. 11211).* 3–19.

[44] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3441–3450.

[45] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4651–4659.

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

[47] Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. 2021. Heterogeneous attention network for effective and efficient cross-modal retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1146–1156.

[48] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018. Rethinking diversified and discriminative proposal generation for visual grounding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 1114–1120.

[49] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. 2020. ResNeSt: Split-attention networks. *arXiv preprint arXiv:2004.08955* (2020).

[50] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. 2020. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3536–3545.