

Two-stream Hierarchical Similarity Reasoning for Image-text Matching

Ran Chen, Hanli Wang, *Senior Member, IEEE*, Lei Wang, Sam Kwong, *Fellow, IEEE*

Abstract—Reasoning-based approaches have demonstrated their powerful ability for the task of image-text matching. In this work, two issues are addressed for image-text matching. First, for reasoning processing, conventional approaches have no ability to find and use multi-level hierarchical similarity information. To solve this problem, a **hierarchical similarity reasoning module** is proposed to automatically extract context information, which is then co-exploited with local interaction information for efficient reasoning. Second, previous approaches only consider learning single-stream similarity alignment (*i.e.*, image-to-text level or text-to-image level), which is inadequate to fully use similarity information for image-text matching. To address this issue, a **two-stream architecture** is developed to decompose image-text matching into image-to-text level and text-to-image level similarity computation. These two issues are investigated by a unifying framework that is trained in an end-to-end manner, namely **two-stream hierarchical similarity reasoning network**. The extensive experiments performed on the two benchmark datasets of MSCOCO and Flickr30K show the superiority of the proposed approach as compared to existing state-of-the-art methods.

Index Terms—Image-text matching, cross-media retrieval, hierarchical similarity reasoning, two-stream network, graph convolutional network.

I. INTRODUCTION

The image-text matching task aims to measure the visual-semantic similarity between image and text, which has a lot of potential applications such as cross-modal retrieval [1]–[3], image captioning [4]–[6], text-to-image synthesis [7], and multi-modal neural machine translation [8]. The traditional methods [9]–[11] constructed common space to maximize the correlation of cross-media information. However, these methods are limited by the hand-crafted features, which require strong prior knowledge. To address this issue, learning based methods have been applied to extract features directly from different modalities. In recent years, research efforts have been devoted to learn the representations of both image and text, and match different modalities based on the learned representations. To this aim, several works [12], [13] mapped the entire image and the full text into a common space,

followed by computing the cosine similarity in the common space. However, these approaches have an insufficiently discriminative ability due to the lack of local interaction between sentence words and image regions. What’s worse, although several techniques [14], [15] have been proposed to extract local interaction, it is difficult to match more complicated image-text pairs by simply grasping these local semantic concepts but overlooking higher-level information.

Recently, higher-level information can be extracted through reasoning-based [16], [17] or attention-based [18] methods. In general, reasoning-based methods firstly adopt different semantics as vertexes to construct an undirected graph. Afterwards, each vertex communicates with its nearest neighbor vertexes. Finally, global representation can be obtained by aggregating all of the vertexes. Specifically, Li *et al* [16] used salient regions as vertexes to reason region relationships for enhancing the original visual features. Unfortunately, limited fine-grained matching is arisen because of unconsidered region-word local interactions. Diao *et al* [17] exploited region-word local similarities as vertexes so that higher-level similarity information was obtained via reasoning these local similarities. However, as shown in Fig. 1(a), the conventional reasoning mechanism tries to realize global representation simply through information communication between directly connected vertexes, thus each vertex ignores to excavate latent information with the non-directly connected vertexes, for instance context information. About attention-based methods, a relation-wise dual attention network (RDAN) was devised in [18] which not only extracted salient objects and key words, but also explored latent relations between objects and words. However, there are two limitations. First, RDAN excavates higher-level information directly based on the cross-attentive map which is misaligned with local similarity, hence the problem of heterogeneity gap cannot be solved and the multilevel information is not accurate enough. Second, due to the lack of reasoning mechanism, RDAN is hard to effectively calculate global similarity.

In general, image-text interaction should be of hierarchy [18]–[20], progressively from local interaction to context-awareness and then to global representation. Concretely, local interaction usually attends to sentence words about image regions; context-awareness attempts to capture several latent information in a local receptive field, such as relative position information in a local space; global representation is more robust by aggregating multi-level information and allowing information communication. However, it is still hard to automatically find and fully utilize hierarchical information without the guidance of external auxiliary knowledge, such as knowledge

Corresponding author: Hanli Wang.

R. Chen and H. Wang are with the Department of Computer Science & Technology, Key Laboratory of Embedded System and Service Computing (Ministry of Education), Tongji University, Shanghai 200092, P. R. China, and with Frontiers Science Center for Intelligent Autonomous Systems, Shanghai 201210, P. R. China (e-mail: 2110142@tongji.edu.cn, hanliwang@tongji.edu.cn).

Lei Wang is with DeepBlue Academy of Sciences, Shanghai 200336, P. R. China (e-mail: wangl@deepblueai.com).

Sam Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong, P. R. China (e-mail: cssamk@cityu.edu.hk).

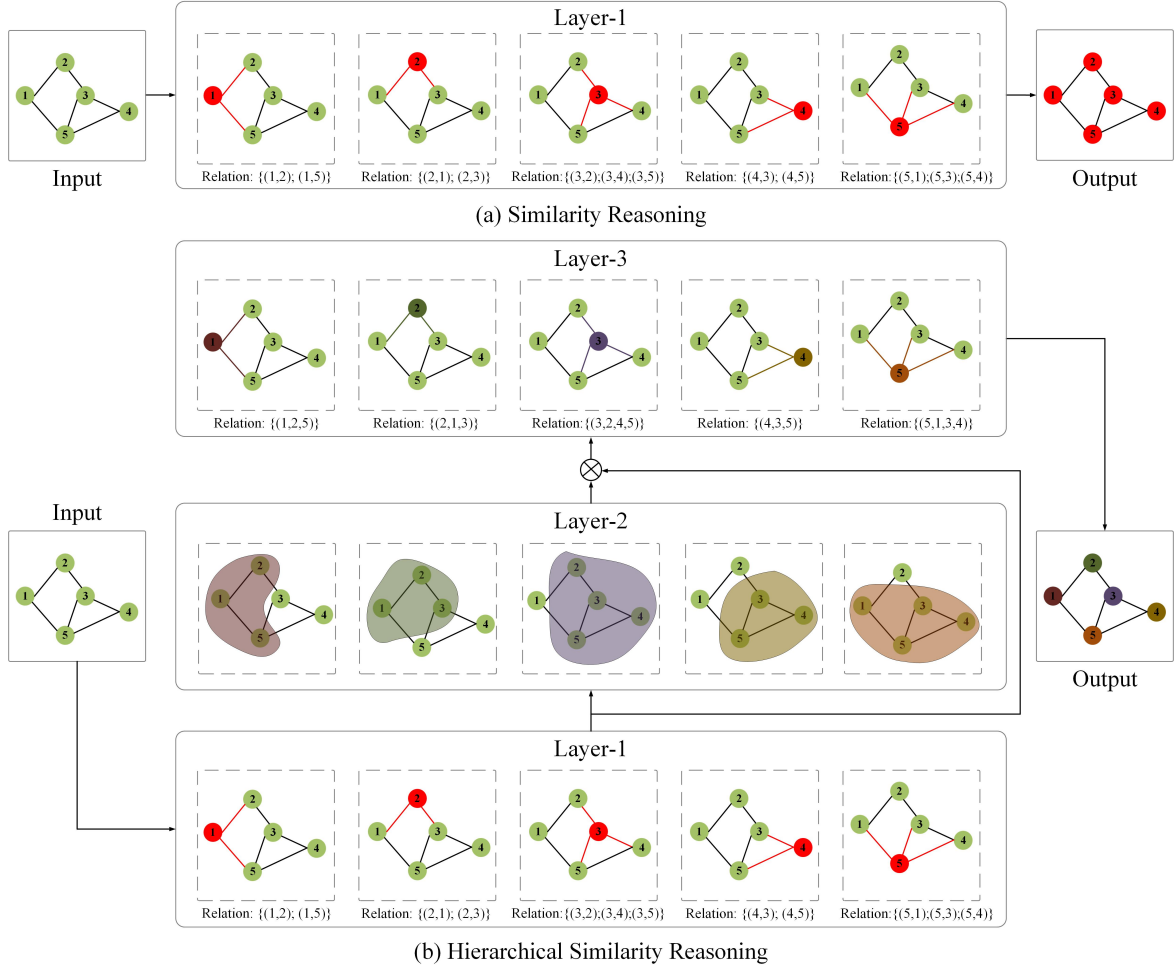


Fig. 1. Illustration of the difference between similarity reasoning [17] and the proposed hierarchical similarity reasoning. Both of these two reasoning strategies start with propagating nearest neighbor information, then (a) the similarity reasoning strategy forms global similarity representations based on the propagated nearest neighbor information directly and ignores the multi-level characteristic of interaction structure, while (b) the proposed hierarchical similarity reasoning strategy automatically finds higher-level information based on the propagated nearest neighbor signals and investigates multi-level signals to form the global information gradually.

mapping [21] and syntactic analysis [22]. Moreover, as shown in Fig. 2, the similarities between image-to-text and text-to-image calculations are different and complementary. Most existing works such as [16]–[18] computed similarity only by adopting image features as cues to attend on texts (image-to-text similarity) or vice versa (text-to-image similarity), where the similarity of image-to-text and text-to-image is not considered in a unified network.

In order to explore hierarchical information sufficiently for the image-text matching task, a **two-stream hierarchical similarity reasoning (TSHSR) network** is designed in this work. In detail, a **hierarchical similarity reasoning (HSR)** module is proposed to fully utilize hierarchical information to progressively form a more robust similarity representation. As shown in Fig. 1(b), there are three layers with different functions in HSR. In the first layer, HSR starts with computing region-word local similarities, and takes these similarities as vertexes, then each vertex exchanges information with its nearest neighbors. In the second layer, to explore context information, an **assembling operation** is adopted to comprehensively

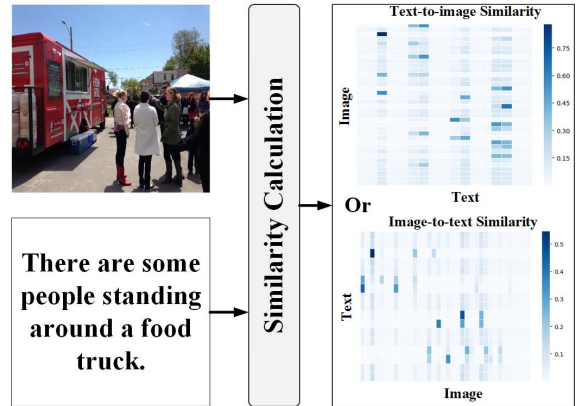


Fig. 2. Illustration of similarity gulf between image-to-text and text-to-image calculations.

investigate each vertex and its nearest neighbors as a whole, as well as fuse lower-level information from the first layer via the operation of **element-wise multiplication**. Consequently, each

vertex collects the information from its nearest neighbors while attending to context information as well. In the third layer, the vertices own more abundant information which boosts more proper reasoning to satisfy sophisticated matching patterns. What's more, both image-to-text similarity and text-to-image similarity are utilized to enrich and supplement neglected global representations, so that global representations can have a stronger representative ability. Moreover, TSHSR is an end-to-end trainable framework.

There are three major contributions of this work. First, a HSR module is proposed for hierarchical similarity reasoning. Different from other reasoning architectures which suffer from the problem of insufficient hierarchical information, HSR enables to automatically recognize multi-level information, which is exploited in an efficient integration manner to achieve more robust global representations. Second, both the image-to-text similarity and text-to-image similarity are utilized to enhance image-text mutual retrieval, since these two similarities are complemented to each other. Third, a TSHSR framework is designed to fully explore the usage of the aforementioned two similarities with a hierarchical similarity reasoning strategy, and an efficient and effective way is devised for TSHSR to optimize its modules in an end-to-end manner. The rest of this paper is organized below. Section II reviews the related works. Section III describes the algorithm of the proposed TSHSR, and Section IV presents experiments to validate the effectiveness of TSHSR. Finally, Section V concludes this work.

II. RELATED WORK

In this section, feature encoding, image-text matching and graph reasoning are reviewed.

A. Feature Encoding

Feature encoding plays an important role for image-text matching. For text encoding, conventional methods [23], [24] tried to utilize SkipGram [25] or Fisher vectors [26] to encode textual features. However, these methods were unable to capture sequential information. To address this issue, Kiros *et al* [27] adopted a GRU [28] instead of SkipGram as text encoder. Regarding image encoding, it can be roughly divided into two classes: coarse-grained representation and fine-grained representation. The efforts of coarse-grained representation try to encode more robust global image representations without resorting to salient features, *e.g.*, Liu *et al* [29] proposed a recurrent residual network that could refine global embeddings, Song *et al* [30] and Wei *et al* [31] both combined global context with locally-guided features employing multi-head self-attention. Besides, some works [32], [33] paid attention to gather semantics by exploiting block-based visual attention on feature maps. Differently, the fine-grained representation uses local salient features for more detailed learning, with the motivation inspired by bottom-up attention [4]. To this aim, several methods [1], [14]–[17], [34] employed object detectors (*e.g.*, faster-RCNN [35]) pre-trained on a large-scale dataset to capture region-based features of visual objects. And then, various algorithms are proposed

based on the captured region-based features to calculate global representation, *e.g.*, Chen *et al* [34] exploited a BiGRU to obtain high-level semantic features, Li *et al* [16] introduced a visual reasoning mechanism to build the relationship between fine-grained visual features. Instead of reasoning on visual features, which has no interaction with textual features, Diao *et al* [17] concentrated on establishing the relationship between constructed visual-textual local similarities. Inspired by this, local similarity representation is employed in this work to progressively form hierarchical interaction, which is different from using local similarity representation directly in [17].

B. Image-Text Matching

After feature encoding, it is required to calculate the final similarity across modalities according to a certain alignment mechanism for image-text matching. Existing alignment methods could be classified to two categories: global alignment [12], [13], [16], [29], [30], [32] and local alignment [14], [18], [36]–[38]. About global alignment, it usually constructs a joint space, and then conducts similarity computation between global image-text representations in this space. Besides, the methods [39], [40] focused on measuring antisymmetric visual-semantic hierarchy by introducing an ordered representation instead of directly computing similarity. As for local alignment, typical methods prefer to compute all region-word pairs and fuse all possible pairs. In general, global-alignment methods have the ability to capture global semantic information while local-alignment methods tend to grasp regional information. By integrating local alignment and global alignment, Diao *et al* [17] proposed to aggregate region-word local similarities. However, existing methods only consider alignment in the image-to-text level, but ignore the text-to-image level alignment which is also important for cross-media retrieval. In this work, both image-to-text level and text-to-image level alignments are applied to enrich cross-attentive information for mutual retrieval.

C. Graph Reasoning

It is beneficial to make information communicated between extracted local regions for enhancing global representation. To address the problem of insufficient communication of information, several graph reasoning approaches [41]–[43] have been extensively used for cross-media computing, such as image captioning [44] and grounding referring expressions [37]. For image-text matching, Shi *et al* [21] aimed to construct scene concept graph with image scene graphs and co-occurring concept pairs, Li *et al* [16] introduced graph convolutional network to build up relation between image regions, Wang *et al* [1] adopted textual and visual scene graphs to refine textual and visual features, Wen *et al* [45] employed graph attention [46] on both visual and textual features for semantic relation reasoning. However, these works have no ability to establish more complex matching patterns because hierarchical information is neglected. In this work, a novel graph reasoning mechanism is designed to effectively excavate hierarchical information for cross-media interaction.

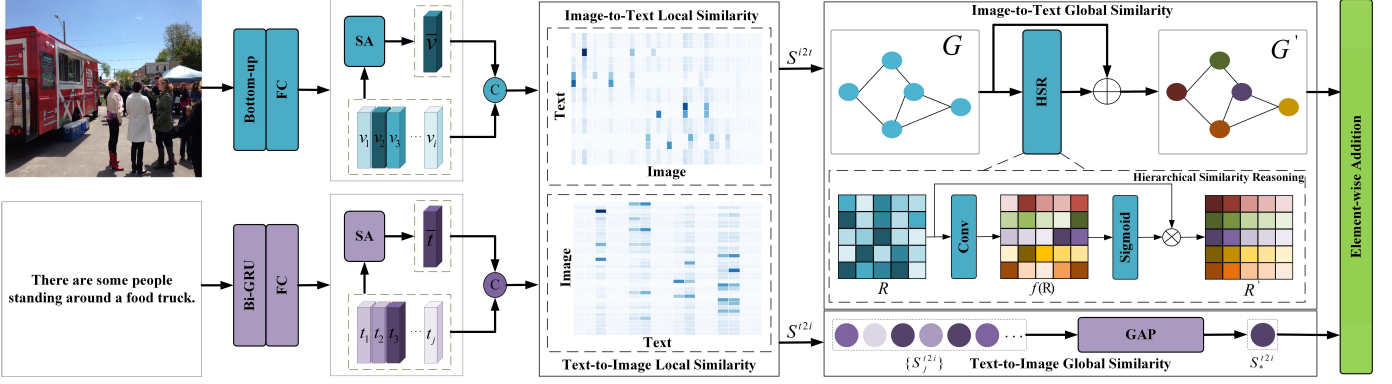


Fig. 3. An overview of the proposed TSHSR framework. For image-to-text similarity, after aligning image-text local features, the HSR module is utilized for global image-to-text similarity representation. For text-to-image similarity representation on the other hand, visual and textual local features are aligned firstly, and then global average pooling is used for global text-to-image similarity representation. An element-wise addition operation is applied to summarize these two global representations.

III. PROPOSED TWO-STREAM HIERARCHICAL SIMILARITY REASONING FOR IMAGE-TEXT MATCHING

The framework of the proposed TSHSR is illustrated in Fig. 3. Technically, for image-to-text similarity representation, aligning image-text local features is firstly started by attending to the sentence word about each image region. Afterwards, the HSR module is used to calculate the global image-to-text similarity. Simultaneously, the text-image local features are firstly aligned via attending to the regions in the sentence with respect to each sentence word, then global average pooling is applied to achieve global text-to-image representation. In the end, these two global representations are summed up to obtain the final similarity representation.

A. Feature Encoding

1) *Image Feature*: For each input image, K region-level visual features [4] are extracted firstly, and then a linear transformation is employed to embed them into d -dimensional vectors as local region representations $V = \{v_1, \dots, v_K\}$, with $v_i \in R^d$. Inspired by [17], the global visual feature $\bar{v} = V \odot \bar{q}_v$ is computed, where $\bar{q}_v = \frac{1}{K} \sum_{i=1}^K v_i$ is calculated to make up for the shortcoming of lacking global embedding, the operator \odot represents element-wise multiplication.

2) *Text Feature*: Given a sentence, it is firstly split into L word tokens with tokenization technique followed by previous works [14], [16], [17], and then the tokens are fed into a BiGRU [47] sequentially to obtain the representation $T = \{t_1, \dots, t_L\}$, with $t_j \in R^d$. Similar to image embedding, global text embedding are also computed by $\bar{t} = T \odot \bar{q}_t$, where $\bar{q}_t = \frac{1}{L} \sum_{j=1}^L t_j$.

B. Local Similarity Representation

Let $S(x, y)$ denote the similarity calculation between x and y . Compared with previous works [14], [16] that immediately used scalar to calculate cosine or Euclidean distance, Diao *et al* [17] used a vector to calculate the distance between modalities to preserve more detailed information by

$$S(x, y) = \frac{W \|x - y\|_2^2}{\|x - y\|_2}, \quad (1)$$

where $W = R^{m \times d}$ is a learnable parameter matrix for obtaining m -dimensional similarity vector, while $|\cdot|^2$ and $\|\cdot\|_2$ indicate element-wise square and l_2 -norm, respectively. Therefore, the similarity between \bar{v} and \bar{t} is computed as $s^g = S(\bar{v}, \bar{t})$.

When local regions are guided to focus on the attended words, the following similarity representation $s^{i2t} = S(\alpha_j^v, t_j)$ is obtained, where $\alpha_j^v = \sum_{i=1}^K \alpha_{ij} v_i$ is the attended visual feature with respect to j -th word, and α_{ij} is the attention weight between region feature v_i and word feature t_j . α_{ij} is computed by

$$\alpha_{ij} = \frac{E(\lambda \bar{c}_{ij})}{\sum_{i=1}^K E(\lambda \bar{c}_{ij})}, \bar{c}_{ij} = \frac{[c_{ij}]_+}{\sqrt{\sum_{j=1}^L [c_{ij}]_+^2}}, \quad (2)$$

where λ is a temperature parameter, $[c_{ij}]_+ = \max(0, \cos(v_i, t_j))$, $\cos(\cdot)$ stands for cosine distance. Similarly, the text-to-image similarity $s^{t2i} = S(v_i, \beta_i^t)$ is generated, where $\beta_i^t = \sum_{j=1}^L \beta_{ij} t_j$ is the attended textual feature with respect to i -th region. β_{ij} is the attention weight between region feature and word feature, and it is computed by

$$\beta_{ij} = \frac{E(\lambda \bar{c}_{ij})}{\sum_{j=1}^L E(\lambda \bar{c}_{ij})}, \bar{c}_{ij} = \frac{[c_{ij}]_+}{\sqrt{\sum_{i=1}^K [c_{ij}]_+^2}}. \quad (3)$$

C. Global Similarity Representation

After obtaining both the image-to-text and text-to-image local similarity representations, the next step is to learn a global similarity representation for image-text matching. To achieve this, s^{i2t} and s^{t2i} are firstly concatenated to form an **initial global feature** $S^{i2t} = \{S_1^{i2t}, \dots, S_L^{i2t}, S_{L+1}^{i2t}\}$, then the proposed HSR is used for hierarchical similarity reasoning. As shown in Fig. 3, a similarity relationship is established between $S_p^{i2t} \in S^{i2t}$ and $S_q^{i2t} \in S^{i2t}$ as

$$\mathcal{R}(S_p^{i2t}, S_q^{i2t}) = \varphi(S_p^{i2t})^T \phi(S_q^{i2t}), \quad (4)$$

where $\varphi(S_p^{i2t}) = W_p S_p^{i2t}$ and $\phi(S_q^{i2t}) = W_q S_q^{i2t}$ are two embeddings, W_p and W_q are learnable parameters.

Afterwards, a similarity relationship graph $G = (\mathcal{S}; \mathcal{R})$ is constructed, where \mathcal{S} is the set of calculated local similarity

representations described by S^{i2t} and \mathcal{R} is the **relation matrix** obtained by calculating the similarity relationship in Eq. (4). Furthermore, a convolution operation is used to assemble n -gram and n -term information in a local receptive field for context-awareness. n -gram occurs with n successive words, and n -term allows for order or semantic alternatives [20]. Finally, \mathcal{R} is re-weighted by the output of convolution followed by a sigmoid activation for multi-level information fusing:

$$\mathcal{R}' = \mathcal{R} \odot \text{sigmoid}(f(\mathcal{R})), \quad (5)$$

where $f(\cdot)$ is a learned convolution kernel with the size of 3×3 in this work. After obtaining the multi-level relation matrix \mathcal{R}' , the response of each similarity representation is updated according to its neighbors that are fused by hierarchical information. Besides, a residual connection is introduced to prevent gradient vanish. Therefore, the image-to-text level global alignment \mathcal{S}_{i2t}^* can be obtained by

$$\mathcal{S}_{i2t}^* = W_r(\mathcal{R}' SW_g) + \mathcal{S}, \quad (6)$$

where W_r and W_g are learnable weight matrices. The number of HSR layers is set as M ($M \geq 1$), and the output of the previous layer is used as the input of the next layer to iteratively reason hierarchical similarity. The output of the global alignment at the last HSR layer is used as the final image-to-text similarity representation.

Regarding text-to-image level global alignment on the other hand, s^{t2i} and s^g are firstly concatenated to form an initial global feature $S^{t2i} = \{S_1^{t2i}, \dots, S_K^{t2i}, S_{K+1}^{t2i}\}$. Different from image-to-text global alignment, a simple global average pooling operation is applied on S^{t2i} to generate the text-to-image level global alignment \mathcal{S}_{t2i}^* as

$$\mathcal{S}_{t2i}^* = \frac{1}{K+1} \sum_{j=1}^{K+1} S_j^{t2i}. \quad (7)$$

At last, an element-wise addition operation is performed on \mathcal{S}_{i2t}^* and \mathcal{S}_{t2i}^* to integrate image-to-text and text-to-image level information to generate the final global similarity representation \mathcal{S}^* as

$$\mathcal{S}^* = \mathcal{S}_{i2t}^* \oplus \mathcal{S}_{t2i}^*. \quad (8)$$

D. Loss Calculation

Following [13], the bidirectional ranking loss is utilized to train the proposed TSHSR. A matched image-text pair $(v; t)$ is given within a mini-batch, the corresponding hardest negative image is denoted as v^- , and the hardest negative text is t^- . The bidirectional ranking loss is computed as

$$L = [\alpha - S_\gamma(v, t) + S_\gamma(v, t^-)]_+ + [\alpha - S_\gamma(v, t) + S_\gamma(v^-, t)]_+, \quad (9)$$

where α serves as a margin parameter, $[x]_+ = \max(x, 0)$, $S_\gamma(\cdot)$ is the similarity function in the joint embedding space, *e.g.*, inner product used in this work.

IV. EXPERIMENT

To verify the effectiveness of the proposed TSHSR approach, a large number of experiments are carried out on two benchmark datasets including MSCOCO [48] and Flickr30K [49]. Section IV-A introduces the datasets, evaluation metrics and implementation details. After that, Section IV-B compares the proposed TSHSR with state-of-the-art methods. Furthermore, ablation study is performed to further assess the configurations of TSHSR in Section IV-C. Finally, several visualization results are exhibited in Section IV-D.

A. Dataset and Setting

1) *Dataset*: The MSCOCO dataset [48] includes 123,287 images, and each image is annotated with 5 annotated captions. In this work, 113,287 images are chosen as the training set, 5000 images as the validation set and 5000 images as the testing set. The results are tested either by averaging over 5 folds of 1K test images or on the full 5K images. The Flickr30K dataset [49] contains 31,783 images, and each image is also annotated with 5 captions. Following [23], 29,783 images are split as the training set, 1000 images as the validation set and the rest as the testing set.

2) *Evaluation Metric*: In this work, the image-text matching performance is measured by recall at K ($R@K$), where the fraction of queries in which the correct item is retrieved in the closest K points to the query.

3) *Implementation Detail*: For image, the Faster-RCNN [35] detector is employed with ResNet-101 [4] to extract the top $K = 36$ salient regions, and each region proposal is encoded to a 2048-dimensional vector. For text, the word embedding size is set as 300, and the size of hidden states is 1024. The dimension of similarity representation is set as 256, with the temperature parameter $\lambda = 9$, the number of hierarchical similarity reasoning steps $M = 3$, and the margin parameter $\alpha = 0.2$. The Adam optimizer [50] is employed to train TSHSR with the mini-batch size of 128 in an end-to-end manner. 20 epochs are trained on MSCOCO, the learning rate is set to be 0.0002 for the first 10 epochs initially and decays by 0.1 for the next 10 epochs. As for Flickr30K, 40 epochs are trained. For the first 30 epochs, the initial learning rate is also 0.0002, and the learning rate is decayed by 0.1 in the last 10 epochs. Snapshot is selected with the best performance used by the summation of the recalls on the validation set.

B. Comparison with State-of-the-Art Methods

1) *Results on MSCOCO 1K and Flickr30K*: Table I shows the quantitative results on MSCOCO 1K and Flickr30K, where it can be seen that the proposed TSHSR outperforms all of the state-of-the-art methods with a large gap on $R@1$. Concretely, TSHSR has the best $R@1=79.0\%$ for sentence retrieval and $R@1=63.1\%$ for image retrieval on MSCOCO 1K. As for Flickr30K, TSHSR also achieves the best on $R@1$, with the best $R@1=76.3\%$ for sentence retrieval and $R@1=56.6\%$ for image retrieval. In addition, when the HSR module is used alone for image-to-text level similarity representation,

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON MSCOCO 1K AND FLICKR30K. BEST PERFORMANCE IS INDICATED BY THE BOLD AND RUNNER-UP BY THE UNDERLINE.

Methods	MSCOCO 1K						Flickr30K					
	Sentence Retrieval			Image Retrieval			Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CAMP [15]	72.3	94.8	98.3	58.5	87.9	95.0	68.1	89.7	95.2	51.5	77.1	85.3
SCAN [14]	72.7	94.8	98.4	58.8	88.4	94.8	67.4	90.3	95.8	48.6	77.7	85.2
SGM [1]	73.4	93.8	97.8	57.5	87.3	94.3	71.8	91.7	95.5	53.5	79.6	86.5
RDAN [18]	74.6	<u>96.2</u>	<u>98.7</u>	61.6	89.2	94.7	68.1	91.0	95.9	54.1	80.9	87.2
MMCA [31]	74.8	95.6	97.7	61.6	<u>89.8</u>	95.2	74.2	92.8	96.4	54.8	<u>81.4</u>	87.8
BFAN [51]	74.9	95.2	-	59.4	88.4	-	68.1	91.4	-	50.8	78.4	-
CAAN [52]	75.5	95.4	98.5	61.3	89.7	95.2	70.1	91.6	97.2	52.8	79.0	<u>87.9</u>
DPRNN [34]	75.3	95.8	98.6	62.5	89.7	95.1	70.2	91.6	95.8	55.5	81.3	88.2
PFAN [53]	76.5	96.3	99.0	61.6	89.6	95.2	70.0	91.8	95.0	50.4	78.7	86.1
VSRN [16]	76.2	94.8	98.2	<u>62.8</u>	89.7	95.1	71.3	90.6	96.0	54.7	81.8	88.2
IMRAM [38]	76.7	95.6	98.5	61.7	89.1	95.0	74.1	<u>93.0</u>	<u>96.6</u>	53.9	79.4	87.2
SGR [17]	78.0	95.8	98.2	61.4	89.3	<u>95.4</u>	75.2	93.3	<u>96.6</u>	<u>56.2</u>	81.0	86.5
Ours (HSR)	<u>78.8</u>	96.3	98.6	<u>62.8</u>	89.9	95.6	<u>76.0</u>	<u>93.0</u>	96.3	55.7	81.3	86.3
Ours (TSHSR)	79.0	<u>96.2</u>	98.6	63.1	89.9	<u>95.4</u>	76.3	<u>93.0</u>	95.8	56.6	81.2	85.9

namely “Ours (HSR)” in Table I, it also performs well, with $R@1=78.8\%$ for sentence retrieval and $R@1=62.8\%$ for image retrieval on MSCOCO 1K, $R@1=76.0\%$ for sentence retrieval and $R@1=55.7\%$ for image retrieval on Flickr30K. According to the results, two conclusions can be obtained. First, HSR improves the capability of image-text matching, which verifies the importance of utilizing hierarchical interaction patterns. Second, when image-to-text level and text-to-image level similarity representations are both considered, namely “Ours (TSHSR)” in Table I, it not only improves the performance of image retrieval, but also boosts the performance of sentence retrieval, which indicates that image-to-text level and text-to-image level similarity representations are complementary.

2) *Results on MSCOCO 5K*: Table II shows the quantitative results on MSCOCO 5K. It is noteworthy that TSHSR also outperforms all of the state-of-the-art methods with about 1% improvement on $R@1$. Besides, HSR also obtains a competitive retrieval performance among almost all of the state-of-the-art methods, demonstrating the necessity and effectiveness of hierarchical similarity reasoning.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON MSCOCO 5K. BEST PERFORMANCE IS INDICATED BY THE BOLD AND RUNNER-UP BY THE UNDERLINE.

Methods	MSCOCO 5K			
	Sentence Retrieval		Image Retrieval	
	R@1	R@10	R@1	R@10
SGM [1]	50.0	87.9	35.3	76.5
CAMP [15]	50.1	89.7	39.0	80.2
SCAN [14]	50.4	90.0	38.6	80.4
CAAN [52]	52.5	90.9	<u>41.2</u>	82.9
VSRN [16]	53.0	89.4	40.5	81.1
IMRAM [38]	53.7	<u>91.0</u>	39.7	79.8
MMCA [31]	54.0	90.7	38.7	80.8
SGR [17]	<u>56.9</u>	90.5	40.2	79.8
Ours (HSR)	56.6	91.4	40.8	80.9
Ours (TSHGR)	57.4	91.4	41.4	<u>81.2</u>

C. Ablation Study

To further explore the impact of different configurations about the proposed TSHSR, ablation studies are conducted on the MSCOCO 1K dataset.

TABLE III
COMPARISON WITH DIFFERENT NUMBER OF HSR LAYERS.

Model	Number				Sentence Retrieval		Image Retrieval	
	0	1	2	3	R@1	R@10	R@1	R@10
1	✓				77.8	98.5	62.0	95.3
2		✓			78.2	98.5	62.2	95.4
3			✓		78.7	98.5	62.6	95.5
4				✓	78.8	98.6	62.8	95.6

1) *How effective is the number of reasoning steps using HSR*: To explore the impact of the number of HSR layers, four numbers are tried as listed in Table III. From the results, it can be seen that with the increase of HSR layers, the performances of both sentence retrieval and image retrieval are improved gradually. This reveals that HSR benefits to both image retrieval and sentence retrieval.

TABLE IV
COMPARISON OF DIFFERENT CONFIGURATIONS OF SIMILARITY REASONING. IN THE FIRST ROW, ‘0’ INDICATES TRADITIONAL SIMILARITY REASONING, AND ‘1’ INDICATES THE PROPOSED HIERARCHICAL SIMILARITY REASONING.

Steps	Config.	0	1	Sen. Ret.		Img. Ret.	
				R@1	R@10	R@1	R@10
1	✓			78.1	98.5	62.4	95.4
			✓	78.2	98.5	62.2	95.4
3	✓			78.3	98.3	62.8	95.5
			✓	78.8	98.6	62.8	95.6

2) *How necessary is hierarchical information*: To explore how important is the hierarchical information for reasoning, several experiments are designed by comparing the performance between hierarchical similarity reasoning and traditional similarity reasoning with different number of reasoning steps. The results are shown in Table IV, where it can be observed that the performances of using hierarchical information

TABLE V
COMPARISON WITH DIFFERENT LEVEL REPRESENTATIONS WHEN USING HSR. I2T INDICATES REASONING ON IMAGE-TO-TEXT LEVEL SIMILARITY, AND T2I INDICATES REASONING ON TEXT-TO-IMAGE LEVEL SIMILARITY.

Alignment \ Steps	Steps			Sen. Ret.		Img. Ret.		Sum
	1	2	3	R@1	R@10	R@1	R@10	
I2T	✓			78.2	98.5	62.2	95.4	334.3
		✓		78.7	98.5	62.6	95.5	335.3
			✓	78.8	98.6	62.8	95.6	335.8
T2I	✓			78.1	98.6	61.6	95.2	333.5
		✓		77.7	98.2	61.5	95.3	332.7
			✓	79.0	98.4	62.0	95.4	334.8


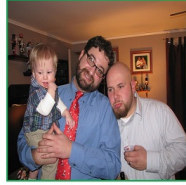







Sentence Retrieval		Image Retrieval		
Query	Retrieved	Query	Retrieved	
	<p>Top 1: Two men and a baby at a Christmas part.</p> <p>Top 2: Two guys making silly faces for a picture.</p> <p>Top 3: Two men and a baby are posing in a room.</p>	Two men and a baby at a Christmas part.		
	<p>Top 1: Elephants move about an enclosed habitat surrounded by a fence.</p> <p>Top 2: Split photo showing elephants in zoo like enclosures.</p> <p>Top 3: An elephant inside of their enclosure playing with logs and eating.</p>	Elephants move about an enclosed habitat surrounded by a fence.		
	<p>Top 1: A person on a snowboard in the snow.</p> <p>Top 2: A couple of horses standing in a field.</p> <p>Top 3: A person stands on a snowboard on a slope.</p>	A person on a snowboard in the snow.		

Fig. 4. Visualization of image-text mutual retrieval by TSHSR on MSCOCO. The top-3 ranked text captions are shown for each image query, and the top-3 ranked images are also illustrated for given text query. Correct matched samples are framed by green boxes while incorrect matched samples are framed by red boxes.

are superior to that without using, demonstrating the necessity of hierarchical information for reasoning.

3) *How important are different level representations:* To explore the importance of different level similarity contributed to image-text matching, an ablation study is conducted with the comparative results shown in Table V. The results from Table V show that reasoning on image-to-text level similarity can obtain better performance than text-to-image level similarity. This is possibly due to the reason that image-to-text level similarity is aligned with natural language information and text-to-image level similarity is aligned with visual information, while natural language information is simpler than visual information. More important, image-to-text level similarity still retains structured information (such as sequential information), which can be extracted easily by convolutional operation. However, the structured information (such as spatial information) about text-to-image level similarity is destroyed.

D. Visualization of Retrieved Results

As shown in Fig. 4, to validate the mutual retrieval performance achieved by TSHSR, the top-3 ranked items are listed according to image query and text query. Specifically, it can be easily found that the top-1 retrieved results are all right not only in sentence retrieval but also in image retrieval, which verifies the superiority of the proposed TSHSR. We also show some incorrect retrieved results in Fig. 4. For example, the top-2 ranked item in the third sentence retrieval is “A couple of horses standing in the field”, which misunderstands the person as a horse. More importantly, TSHSR has the ability to capture complicated matching patterns. For instance, in the first sample of sentence retrieval, TSHSR not only captures the entities of three persons as shown in the top-1 ranked item, but also detects the tiny expression as shown in the top-2 ranked item.

V. CONCLUSION

In order to tackle the two problems of insufficient hierarchical information and inadequate information utilization in the image-text matching task, a simple and effective architecture

TSHSR is presented to enable the reasoning procedure to spontaneously employ hierarchical information and utilize both image-to-text level as well as text-to-image level similarity information. The proposed TSHSR is evaluated on two benchmark datasets, and the comparative results demonstrate the effectiveness of the design principle and the better performance over other state-of-the-art methods. Although current reasoning approaches can construct the relationship between two entities, such as “move”, “play”, “show”, etc, positional information is usually overlooked in the reasoning process, such as “in”, “on”, “over” and so on. In the future, it is desired to explore the importance of positional information for reasoning.

REFERENCES

- [1] S. J. Wang, R. P. Wang, Z. W. Yao, S. G. Shan, and X. L. Chen, “Cross-modal scene graph matching for relationship-aware image-text retrieval,” in *Proc. WACV’20*, Mar. 2020, pp. 1497–1506.
- [2] X. Sun, H. Wang, and B. He, “MABAN: Multi-agent boundary-aware network for natural language moment retrieval,” *IEEE Trans. Image Process.*, vol. 30, pp. 5589–5599, Jun. 2021.
- [3] J. Tang, K. Wang, and L. Shao, “Supervised matrix factorization hashing for cross-modal retrieval,” *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3157–3166, Jul. 2016.
- [4] P. Anderson, X. D. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proc. CVPR’18*, Jun. 2018, pp. 6077–6086.
- [5] L. Yang, H. Wang, P. Tang, and Q. Li, “Captionnet: A tailor-made recurrent neural network for generating image descriptions,” *IEEE Trans. Multimedia*, vol. 23, pp. 835–845, Apr. 2021.
- [6] N. G. Yu, X. L. Hu, B. H. Song, J. Yang, and J. W. Zhang, “Topic-oriented image captioning based on order-embedding,” *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2743–2754, Jun. 2019.
- [7] T. Xu, P. C. Zhang, Q. Y. Huang, H. Zhang, Z. Gan, X. L. Huang, and X. D. He, “Attngan: Fine-grained text to image generation with attentional generative adversarial networks,” in *Proc. CVPR’18*, Jun. 2018, pp. 1316–1324.
- [8] J. Toyama, M. Misono, M. Suzuki, K. Nakayama, and Y. Matsuo, “Neural machine translation with latent semantic of image and text,” in *arXiv:1611.08459*, Nov. 2016.
- [9] J. W. Qi, Y. X. Peng, and Y. X. Yuan, “Cross-media multi-level alignment with relation attention network,” in *Proc. IJCAI’18*, Jul. 2018, pp. 892–898.
- [10] D. R. Hardoon, S. Sandor, and S. T. John, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.
- [11] Y. C. Gong, Q. F. Ke, I. Michael, and L. Svetlana, “A multi-view embedding space for modeling internet images, tags, and their semantics,” *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, Jan. 2014.
- [12] L. W. Wang, Y. Li, and L. Svetlana, “Learning deep structure-preserving image-text embeddings,” in *Proc. CVPR’16*, Jun. 2016, pp. 5005–5013.
- [13] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “VSE++: Improving visual-semantic embeddings with hard negatives,” in *arXiv:1707.05612*, Jul. 2017.
- [14] K. H. Lee, X. Chen, G. Hua, H. D. Hu, and X. D. He, “Stacked cross attention for image-text matching,” in *Proc. ECCV’18*, Sep. 2018, pp. 201–216.
- [15] Z. H. Wang, X. H. Liu, H. S. Li, L. Sheng, J. J. Yan, X. G. Wang, and J. Shao, “CAMP: Cross-modal adaptive message passing for text-image retrieval,” in *Proc. ICCV’19*, Oct. 2019, pp. 5763–5772.
- [16] K. P. Li, Y. L. Zhang, K. Li, Y. Y. Li, and Y. Fu, “Visual semantic reasoning for image-text matching,” in *Proc. ICCV’19*, Oct. 2019, pp. 4653–4661.
- [17] H. W. Diao, Y. Zhang, L. Ma, and H. C. Lu, “Similarity reasoning and filtration for image-text matching,” in *Proc. AAAI’21*, May 2021, pp. 1218–1226.
- [18] Z. B. Hu, Y. S. Luo, J. Lin, Y. Yan, and J. Chen, “Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching,” in *Proc. IJCAI’19*, Aug. 2019, pp. 789–795.
- [19] L. Ma, Z. D. Lu, L. F. Shang, and H. Li, “Multimodal convolutional neural networks for matching image and sentence,” in *Proc. ICCV’15*, Apr. 2015, pp. 2623–2631.
- [20] P. Liang, Y. Lan, J. Guo, J. Xu, and X. Cheng, “Text matching as image recognition,” in *Proc. AAAI’16*, Mar. 2016, pp. 2793–2799.
- [21] B. T. Shi, L. Ji, P. Lu, Z. D. Niu, and N. Duan, “Knowledge aware semantic concept expansion for image-text matching,” in *Proc. IJCAI’19*, Jul. 2019, pp. 5182–5189.
- [22] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, “Cross-modal progressive comprehension for referring segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell. (Early Access)*, May 2021.
- [23] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, “DeViSE: A deep visual-semantic embedding model,” in *Proc. NIPS’13*, Dec. 2013, pp. 2121–2129.
- [24] B. Klein, G. Lev, G. Sadeh, and L. Wolf, “Associating neural word embeddings with deep image representations using fisher vectors,” in *Proc. CVPR’15*, Jun. 2015, pp. 4437–4446.
- [25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proc. ICLR’13 Workshops*, Jan. 2013.
- [26] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in *Proc. CVPR’07*, Jun. 2007, pp. 1–8.
- [27] R. Kiros, R. Salakhutdinov, and R. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” in *Proc. ICML’14*, Nov. 2014, pp. 595–603.
- [28] J. Y. Chung, C. Gulcehre, K. Y. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in *Proc. NIPS’14 Workshops*, Dec. 2014, pp. 1–9.
- [29] Y. Liu, Y. M. Guo, E. M. Bakker, and M. S. Lew, “Learning a recurrent residual fusion network for multimodal matching,” in *Proc. ICCV’17*, Oct. 2017, pp. 4127–4136.
- [30] Y. L. Song and M. Soleymani, “Polysemous visual-semantic embedding for cross-modal retrieval,” in *Proc. CVPR’19*, Jun. 2019, pp. 1979–1988.
- [31] X. Wei, T. Z. Zhang, Y. Li, Y. D. Zhang, and F. Wu, “Multi-modality cross attention network for image and sentence matching,” in *Proc. CVPR’20*, Jun. 2020, pp. 10938–10947.
- [32] H. Nam, J. W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in *Proc. CVPR’17*, Jul. 2017, pp. 2156–2164.
- [33] Z. Ji, H. R. Wang, J. G. Han, and Y. W. Pang, “Saliency-guided attention network for image-sentence matching,” in *Proc. ICCV’19*, Oct. 2019, pp. 5753–5762.
- [34] T. L. Chen and J. B. Luo, “Expressing objects just like words: Recurrent visual embedding for image-text matching,” in *Proc. AAAI’20*, Jul. 2020, pp. 10583–10590.
- [35] S. Q. Ren, K. M. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Proc. NIPS’15*, Jun. 2015, pp. 91–99.
- [36] A. Karpathy and F. F. Li, “Deep visual-semantic alignments for generating image descriptions,” in *Proc. CVPR’15*, Jun. 2015, pp. 3128–3137.
- [37] P. Wang, Q. Wu, J. W. Cao, C. H. Shen, L. L. Gao, and A. van den Hengel, “Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks,” in *Proc. CVPR’19*, Jun. 2019, pp. 1960–1968.
- [38] H. Chen, G. G. Ding, X. D. Liu, Z. J. Lin, J. Liu, and H. G. Han, “IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval,” in *Proc. CVPR’20*, Jun. 2020, pp. 12655–12663.
- [39] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, “Order-embeddings of images and language,” in *Proc. ICLR’16*, Nov. 2016.
- [40] J. X. Gu, J. F. Cai, S. R. Joty, L. Niu, and G. Wang, “Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models,” in *Proc. CVPR’18*, Jun. 2018, pp. 7181–7189.
- [41] D. K. David, M. Dougal, I. Jorge, B. Rafael, H. Timothy, A. Alan, and P. A. Ryan, “Convolutional networks on graphs for learning molecular fingerprints,” in *Proc. NIPS’15*, Sep. 2015, pp. 2224–2232.
- [42] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *arXiv:1609.02907*, Feb. 2017.
- [43] Y. J. Li, Z. Richard, M. Brockschmidt, and D. Tarlow, “Gated graph sequence neural networks,” in *Proc. ICLR’16*, Apr. 2016.
- [44] X. Yang, K. H. Tang, H. W. Zhang, and J. F. Cai, “Auto-encoding scene graphs for image captioning,” in *Proc. CVPR’19*, Jun. 2019, pp. 10677–10686.
- [45] K. Y. Wen, X. D. Gu, and Q. R. Cheng, “Learning dual semantic relations with graph attention for image-text matching,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2866–2879, Jul. 2021.
- [46] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in *Proc. ICLR’18*, Feb. 2018.

- [47] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Dec. 1997.
- [48] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. ECCV'14*, Sep. 2014, pp. 740–755.
- [49] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," in *Proc. TACL'14*, Feb. 2014, pp. 67–78.
- [50] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR'14*, Dec. 2014.
- [51] C. X. Liu, Z. D. Mao, A. A. Liu, T. Z. Zhang, B. Wang, and Y. D. Zhang, "Focus your attention: A bidirectional focal attention network for image-text matching," in *Proc. ACMMM'19*, Oct. 2019, pp. 3–11.
- [52] Q. Zhang, Z. Lei, Z. X. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proc. CVPR'20*, Jun. 2020, pp. 3533–3542.
- [53] Y. X. Wang, H. Yang, X. M. Qian, L. Ma, J. Lu, B. Li, and X. Fan, "Position focused attention network for image-text matching," in *Proc. IJCAI'19*, Aug. 2019, pp. 3792–3798.