Cross-modal Graph Matching Network for Image-text Retrieval

YUHAO CHENG, XIAOGUANG ZHU, JIUCHAO QIAN, FEI WEN, and PEILIN LIU,
Shanghai Jiao Tong University, China

Image-text retrieval is a fundamental cross-modal task whose main idea is to learn image-text matching. Generally, according to whether there exist interactions during the retrieval process, existing image-text retrieval methods can be classified into independent representation matching methods and cross-interaction matching methods. The independent representation matching methods generate the embeddings of images and sentences independently and thus are convenient for retrieval with hand-crafted matching measures (e.g., cosine or Euclidean distance). As to the cross-interaction matching methods, they achieve improvement by introducing the interaction-based networks for inter-relation reasoning, yet suffer the low retrieval efficiency. This article aims to develop a method that takes the advantages of cross-modal inter-relation reasoning of cross-interaction methods while being as efficient as the independent methods. To this end, we propose a graph-based **Cross-modal Graph Matching Network (CGMN)**, which explores both intra- and inter-relations without introducing network interaction. In CGMN, graphs are used for both visual and textual representation to achieve intra-relation reasoning across regions and words, respectively. Furthermore, we propose a novel graph node matching loss to learn fine-grained cross-modal correspondence and to achieve inter-relation reasoning. Experiments on benchmark datasets MS-COCO, Flickr8K, and Flickr30K show that CGMN outperforms state-of-the-art methods in image retrieval. Moreover, CGMM is much more efficient than state-of-the-art methods using interactive matching. The code is available at https://github.com/cyh-sj/CGMN.

CCS Concepts: • **Computing methodologies → Visual content-based indexing and retrieval;**

Additional Key Words and Phrases: Image-text retrieval, relation reasoning, graph matching, cross-modal matching

ACM Reference format:
Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. 2022. Cross-modal Graph Matching Network for Image-text Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 4, Article 95 (February 2022), 23 pages.
https://doi.org/10.1145/3499027

1 INTRODUCTION

Vision and language are two common modalities for humans to obtain and express information. With the development of multimedia technology, it is necessary to construct the connection between vision and language. Many researches have been dedicated to the intersection of

Authors' address: Y. Cheng, X. Zhu, J. Qian, F. Wen (corresponding author), and P. Liu, Shanghai Jiao Tong University, Shanghai, Shanghai, China, 200240; emails: {cyh958859352, Zhuxiaoguang178, jcqian, wenfei, liupeilin}@sjtu.edu.cn.

1551-6857/2022/02-ART95 $15.00
https://doi.org/10.1145/3499027

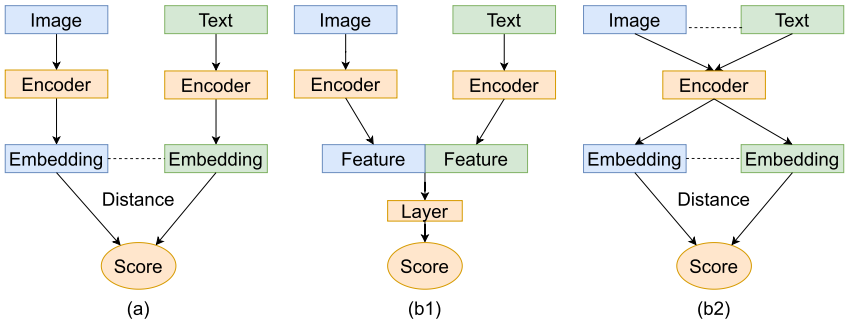

Fig. 1. Illustration of the two different matching methods. (a) Independent representation matching. Cross-interaction matching (b1) without cross-embedding extraction or (b2) with cross-embedding extraction. The difference between them is whether there is an interaction during retrieval process.

computer vision and natural language processing named cross-modal tasks, such as cross-modal retrieval [18, 29], VQA [14, 38], vision language navigation [2, 46], and other cross-modal tasks [30, 31, 35, 54, 57]. These tasks, which attempt to fill the gap between visual and semantic information, have attracted the attention of academia and industry and become a popular research field.

Image-text retrieval is a fundamental cross-modal task in the context of the rapidly increasing of multimedia data from social media and the web. Its primary goal is to find the most similar sentences to an image query or retrieve the closest matching images to a sentence query. Image-text retrieval can be widely used in online shopping, search system, and social network. Generally, according to whether there exist interactions during retrieval process, existing image-text retrieval methods can be classified into two categories, including independent representation matching methods and cross-interaction matching methods, as illustrated in Figure 1.

Typically, independent representation matching methods encode images and texts into a common subspace separately without any cross-modality interaction (Figure 1(a)). Then, the cross-modality similarity is computed from the representation based on hand-crafted measures, such as Euclidean or cosine distance [11, 18]. These methods use global embeddings to represent images and texts using a deep network with two branches, whose main idea is to find an approach to jointly represent images and text [10, 17, 23, 44]. A significant character of cross-interaction matching methods is the adoption of an interactive process between the two modalities to better learn cross-modal correspondence (Figure 1(b1) and Figure 1(b2)). Such methods can be further divided into two classes in terms of whether to generate final embeddings for the two modalities. One is to use a network for similarity measure [6, 7, 48, 49] (Figure 1(b1)), and the other is to generate embeddings via interaction between the two modalities such as inter-modal attention mechanism [4, 5, 22, 56] (Figure 1(b2)). Recently, cross-interaction matching methods have demonstrated significant superiority over independent representation matching methods in terms of matching accuracy.

However, there is a dilemma of retrieval efficiency and accuracy between these two methods. Local region features and text features are adopted for fine-grained correspondence to improve matching accuracy in cross-combined matching methods. However, as the query is processed with all the items in the database by a complex network-based interactive comparison, similarity computing in an interactive manner or network-based matching results in low computational efficiency. In image-text retrieval, an image (respectively, sentence) query is compared with each sentence (respectively, image) in the search database. When using the independent representation matching methods, the matching step only needs to compute the embedding distances between

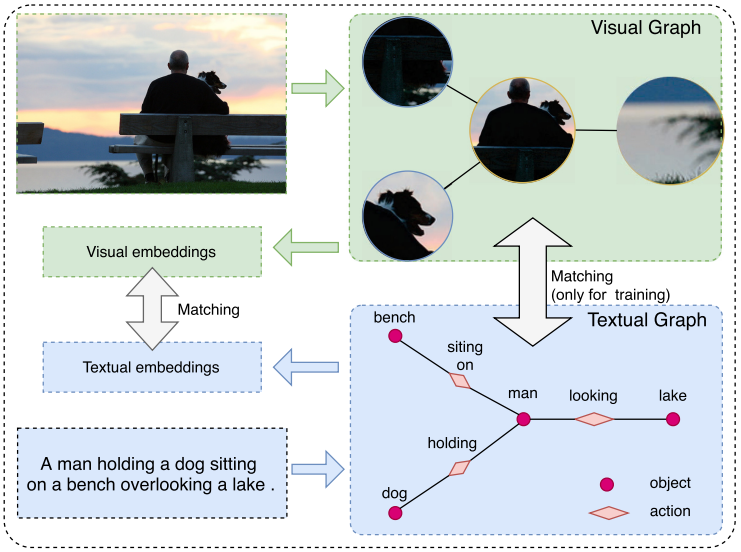

Fig. 2. Illustration of our CGMN in image-text retrieval. Both image and text are represented as graphs and then encoded into embeddings for matching. Graph node matching is used to promote region-word pair alignment to learn the fine-grained cross-modal correspondence and inter-relation reasoning between vision and language during training.

the query and each pre-stored image or sentence embedding in the database. This procedure is far more computationally efficient than cross-combined matching methods. Therefore, independent representation matching methods would be more preferred in image-text retrieval applications due to their computational simplicity, but at some sacrifice of accuracy. Therefore, it is essential to develop an effective and efficient image-text matching method, which can achieve good accuracy as cross-combined matching methods, while being as efficient as independent representation matching methods.

To fulfill the goal mentioned above, we propose a **Cross-modal Graph Matching Network (CGMN)** for fine-grained and fast image-text retrieval, which explores intra-relation in images and sentences, respectively, and achieves inter-relation reasoning between regions and words without affecting search efficiency, as shown in Figure 2. First, we use graphs to represent both images and text, which are constructed by semantic relation and spatial relation. The semantic relation explores the potential connection between image regions or words, and the spatial relation represents the positional relations. Moreover, the combination of them better represents the relations between these cross-modal data. Then, we adopt **Graph Convolution Networks (GCNs)** [20] for intra-relation reasoning among regions and words, respectively, to better jointly represent the two modalities. Second, we propose a novel cross-modal graph node matching method to promote cross-modal alignment and inter-relation reasoning between the visual and textual graphs. The graph node matching is only used for node interaction during training, but not used in network inference for a test. In the training phase, it is used in combination with the metric loss to better optimize the network. Hence, CGMN maintains high efficiency as independent representation matching methods in the test.

To test the effectiveness of our model, we conducted experiments on three popular benchmarks, Flickr8K [55], Flickr30K [55], and MS-COCO [27]. The results show that our proposed CGMN outperforms current SOTA independent representation matching methods and is competitive to

the SOTA cross-interaction matching methods. Moreover, our CGMN is much more efficient than cross-interaction methods. The main contributions of this article are as follows:

1. We propose a novel graph-based independent representation method CGMN for fine-grained and fast image-text retrieval, which is computationally eficient as independent representation methods while taking the advantage of cross-modal inter-relation reasoning of cross-interaction methods.

2. We design a graph-based network to achieve intra-relation reasoning in embedding images and sentences. Particularly, we propose a novel graph node matching loss only used during training, to better learn cross-modal fine-grained alignment and achieve inter-relation reasoning between image regions and words in sentences, without any sacrifice of computational efficiency in the retrieval.

3. In terms of Recall@1 on MS-COCO1K, our model outperforms the current SOTA independent representation matching method VSRN [24] by 0.6% in sentence retrieval and 1.0% in image retrieval, respectively. On Flickr30K, the improvement of Recall@1 is 6.6% in sentence retrieval and 5.2% in image retrieval. Meanwhile, CGMN achieves competitive results as the SOTA cross-interaction matching method GSMN [28]. On Flickr8K [55], CGMN achieves 2.7% and 0.3% improvement over IMRAM [4] on Recall@1 in sentence retrieval and image retrieval, respectively. Besides, our method is far more computationally efficient than cross-interaction matching methods.

2 RELATED WORK

2.1 Independent Representation Matching Methods

Independent representation matching methods learn the approach to jointly represent images and sentences, which focuses on the network structure design of image and sentence processing. In such methods, embeddings of images and sentences are generated by end-to-end networks independently, then the similarity is computed from the representation based on hand-crafted measures.

Most early works on image-text retrieval are independent representation matching methods. In these methods, global features are extracted for the final image-sentence embeddings [13, 41, 44, 60]. DeViSE [13] uses textual data to learn semantic relationships between labels, then embeds the images into a rich semantic embedding space for image-sentence matching. Socher et al. [41] propose an SDT-RNN with constituency trees, which focuses on the action and agents in a sentence. In Reference [44], CNN and RNN are applied to encode images and sentences, respectively, to obtain global embeddings, and then triplet loss is used to learn cross-modal matching. Zheng et al. [60] propose a dual-path CNN for visual-textual embedding learning and a novel instance loss for classification.

Then, some researchers begin to focus on using pre-extracted regional objects and word-level objects for final matching [16, 19]. Karpathy et al. [19] embed fragments of objects of images and fragments of sentences into a common space, then structured max-margin is applied to associate these fragments. Huang et al. [16] employ a classification model on image regions to extract specific objects to improve image embedding.

Recently, it has been recognized that the relationship between different objects can help the embedding and matching of images and texts, and many works have begun to focus on relationship extraction and reasoning [24, 45, 51, 52]. Wu et al. [52] apply the self-attention layer for potential and fine-grained fragment relation reasoning, which also identifies the semantically salient regions. Li et al. [24] introduce a reasoning network that uses GCN for relation reasoning, which shows significant effectiveness in independent representation matching. Wang et al. [45] adopt scene graph

matching to measure the similarity between images and texts. Wen et al. [51] use global features and local features fusion for image embedding, where **Graph Attention Networks (GATs)** [43] are employed for relation reasoning. The sentences are also built as graph structures for relation reasoning.

All the methods above generate the final embeddings via end-to-end models without any interaction. However, without interaction, they only consider the intra-relation reasoning in images or sentences, respectively, while the inter-relation between images and sentences is not considered.

2.2 Cross-interaction Matching Methods

Cross-interaction matching methods use cross-interaction models or functions, i.e., attention mechanism, similarity measure networks, transformers [42], for inter-relation extraction between the images and sentences. It can be divided into interaction embedding generating methods [22, 39, 47, 50, 59] and network-based similarity measure methods [9, 28].

In interaction embedding generating methods, embeddings of images and sentences are generated by networks with cross-modal interaction, and then the similarity is measured by hand-crafted methods. Plummer et al. [39] propose a method that simplifies the representation requirements for individual embeddings and shares the representations before being proposed. Lee et al. [22] represent the images and sentences by region-level features and word-level features and propose a stacked cross attention network for latent image-text alignments. Wang et al. [47] add positional relationships to the region embedding for better matching. Besides, they design a novel positional regions embedding method for better fine-grained matching. Zhang et al. [59] apply cross-attention and self-attention on images and sentences to explore the intra-relations and inter-relations. Wei et al. [50] employ transformers in each modality and between the two modalities for cross-relation and self-relation extraction.

Network-based similarity measure methods use a learnable network to calculate the similarity directly. Liu et al. [28] propose a novel graph matching network with node-level matching and structure-level matching for fine-grained matching. Diao et al. [9] propose a similarity graph reasoning module to learn fine-grained local and global alignments and an attention filtration module for more effective alignment.

Besides, some works focus on image-text pre-trained models with transformers for image-text matching. For example, Lu et al. [32] propose VilBert, which is trained based on transformers, to achieve joint representation of image and text in multiple vision-language tasks. In Reference [25], a saliency detection model is trained to detect the salient parts in the image and uses the labels of these objects as anchor points to learn the alignment of images and texts.

Different from these methods, our work takes advantage of both independent representation methods and cross-interaction matching methods, in which we employ cross-modal alignment with inter-relation reasoning to help fine-grained learning, and similarity is calculated via independent embedding matching to achieve efficient retrieval.

2.3 Graph Matching

The graph structure is a powerful representation form of unstructured information, including information on different nodes and their mutual relationship. Graph matching is a method to calculate the similarity between graphs. Graph matching is an NP-hard problem, which is of high computational complexity. Recent methods use deep learning for approximate graph matching. In Reference [58], the authors first propose a deep learning method for graph matching. Li et al. [26] contribute to generating graph embeddings for similarity learning and propose a new graph matching network that adopts cross-graph attention matching. In Reference [3], Bai et al. propose a simple but fast graph matching method SimGNN, which embeds the graphs and uses a

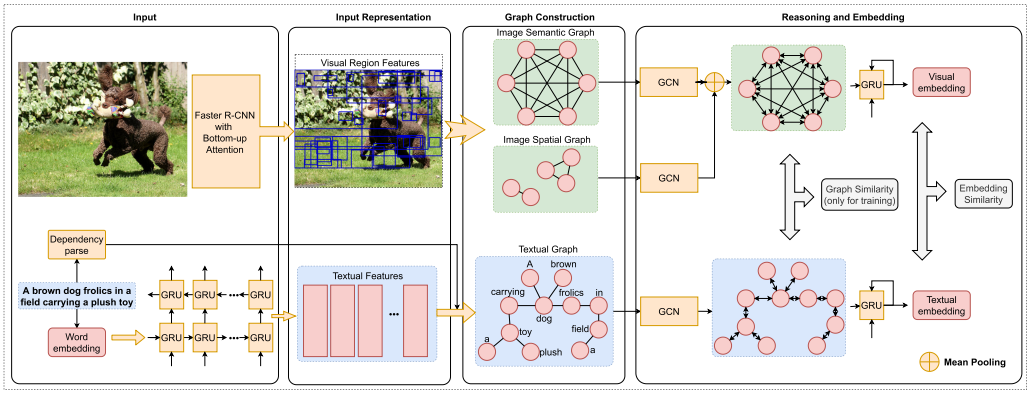

Fig. 3. Theworkflow of our proposed Cross-modal Graph Matching Network (CGMN). (1) The input images are processed by Faster R-CNN with Bottom-up attention to obtain region features, and the sentences are embedded into word vectors and proposed by Stanford CoreNLP for obtaining grammatical dependency. (2) Visual graphs are constructed by semantic relation and spatial relation, while the textual graph is constructed via grammatical dependency and semantic similarity. (3) GCNs are employed to infer the relations between different nodes, then GRUs are used to obtain embedding representation. (4) Graph node similarity and embedding similarity matching are applied as the objective function to train the model. Note that graph node similarity matching is only applied during training to better optimize the network without influencing the efficiency.

neural tensor network to combine the graph embeddings for similarity measurement. Moreover, a two-stage network is proposed to pass node information to the neighborhood for graph node-pair matching in Reference [12]. In knowledge graph alignment, Xu et al. [53] combine node matching with structure matching and have verified the efficacy of graph node matching. Inspired by these methods, graph node matching is employed in visual graphs and textual graphs matching to help learn fine-grained correspondence and inter-relation reasoning between regions in images and words in texts.

3 METHOD

This section elaborates on our proposed CGMN model in detail. The workflow of our model is illustrated in Figure 3. We first introduce the methods to generate image region features and textual features. Afterward, we present the way to construct visual and textual graphs. Then, we describe the relation reasoning network for extracting image and text embeddings. Finally, the training method and the novel graph matching loss are presented.

3.1 Image and Text Representation

Image Representation. Following Reference [22], given an image P, we use Faster R-CNN [40] with bottom-up attention [1] pretrained by Visual Genome [21] to detect m salient regions and obtain their region features $R = \{r_i | i = 1, \ldots, m, r_i \in \mathbb{R}^D\}$. Then, these features are encoded by a fully connected layer into a D-dimensional space:

$$v_i = W_f r_i + b_f, \tag{1}$$

where W_f is the parameter matrix of the fully connected layer and b_f is a bias vector. By this process, the image is represented as $V = \{v_1, v_2, \ldots, v_m\}$.

Text Representation. Given a sentence L with l words represented by one-hot vectors, we first embed them into 300-dimensional feature vectors $E_l = \{e_1, e_2, \ldots, e_l\}$. Then a **Bi-directional Gated Recurrent Union (Bi-GRU)** [8] is applied to encode the sentence from both forward and backward directions:

$$\begin{aligned} \overrightarrow{h_i} &= \overrightarrow{GRU}(e_i, \overrightarrow{h_{i-1}}), \quad i \in \{1, 2, \ldots, l\}, \\ \overleftarrow{h_i} &= \overleftarrow{GRU}(e_i, \overleftarrow{h_{i+1}}), \quad i \in \{1, 2, \ldots, l\}, \end{aligned} \tag{2}$$

where $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ denote the hidden states from the forward and backward directions, respectively. Finally, we obtain the word-level textual features $S = \{s_1, s_2, \ldots, s_l\}$ with s_i given by

$$s_i = \frac{\overrightarrow{h_i} + \overleftarrow{h_i}}{2}, \quad i \in \{1, 2, \ldots, l\}. \tag{3}$$

3.2 Graph Construction

Visual Graph Construction. For intra-relation reasoning between image regions, we design two kinds of graphs to represent the images with potential semantic relation and spatial relation. When describing an image, we generally explore the possible relationships between the regions or entities that appear in the image. For example, when a person and a basketball appear, we may define this action as "playing" or "carrying." Meanwhile, the position distance between different regions or entities may also influence the cognition of their relation. When two objects are very close, they may have a certain connection, and on the contrary, if they are very far, then there may be no connection between them. To represent the spatial connection and potential semantic connection between different regions in the image, we construct two graphs, including a spatial graph $G_{sp} = (V_{sp}, E_{sp})$ and a semantic graph $G_{se} = (V_{se}, E_{se})$. Note that the spatial relations can be calculated directly, while the potential semantic relations need to be learned during training.

The spatial graph $G_{sp} = (V_{sp}, E_{sp})$ is used to represent positional relationship between regions in the image. Compared to the relative distance that cannot accurately distinguish objects of different sizes, **Intersection over Union (IoU)** is used in our CGMN to build positional relations. Intuitively, there is a high probability of a potential relationship between two overlapping regions between overlapping regions. The node-set V_{sp} of G_{sp} is built on the image representation V. The weight W_{sp} of edge E_{sp} depends on the IoU of pair-wise regions. Moreover, the overlapping regions contain similar semantic information that benefits the spatial relation reasoning in the spatial graph. To better construct the correlation between regions, semantic similarity is added to the weight of the spatial graph to help relationship reasoning. Specifically, denoting the IoU between the (i, j)-th region pair by $IoU_{i,j}$, the corresponding weight of this region pair is defined by:

$$W_{sp_{i,j}} = \begin{cases} cos(v_i, v_j) \times IoU_{i,j}, & IoU_{i,j} \geq \xi \\ 0, & IoU_{i,j} < \xi \end{cases}, \tag{4}$$

where $cos(\cdot, \cdot)$ is the cosine function and ξ is the threshold.

The semantic graph $G_{se} = (V_{se}, E_{se})$ represents region features and the potential semantic connection between regions. Similarly, V_{se} is built on region features V, and E_{se} is the edge set of the graph structure expressed by adjacency matrix W_{se}, which represents the connection between regions. Following Reference [24], we use a region relation reasoning model to describe the relationship among image regions as:

$$W_{se_{i,j}} = \varphi(v_i)^T \phi(v_j). \tag{5}$$

$\varphi(v_i) = W_\varphi v_i$ and $\phi(v_j) = W_\phi v_j$ are two feature embeddings, in which W_φ and W_ϕ are parameters to be learned. The semantic graph $G_{se} = (V_{se}, E_{se})$ is a fully connected graph, and a large value of an edge between any two nodes means that there is a strong semantic relation.

Textual Graph Construction. Similar to images, the sentences are also constructed as textual graphs. The textual graph is constructed as $G_t = (V_t, E_t)$ for each sentence where V_t is built on the textual features S with the edge weight given by the matrix W_t added self-loops.

There is a certain interpretable grammatical dependency in a sentence. For example, given a sentence "A black and white dog is running through the grass," "A," "black," and "white" are the attributes of the entity "dog," while "running" is an action and "through the grass" is a condition. A grammatical dependency matrix W_d is produced by Stanford CoreNLP [34]. By the software, $W_{d_{i,j}} = 1$ if there exists grammatical dependency between s_i and s_j, otherwise $W_{d_{i,j}} = 0$. Similar to W_{sp}, the similarity between words helps emphasize the intra-relation in sentences. Therefore, we take the pair-wise similarity between nodes into the weight matrix as:

$$W_{t_{i,j}} = cos(s_i, s_j) \times W_{d_{i,j}}. \tag{6}$$

3.3 Reasoning and Embedding Method

Visual Reasoning and Embedding. Aiming to explore the intra-relation reasoning among regions and words, respectively, we apply the **Graph Convolutional Network (GCN)** [24] on a constructed k-nodes graph $G = (V, E), V \in \mathbb{R}^{k \times D}, W \in \mathbb{R}^{k \times k}$, where D is the dimension of the nodes. GCN is a special model for processing graph structure, which can infer the potential relationship between certain nodes. The value of the edge between the nodes determines the information transferred. The function of graph updating is defined as:

$$V^* = \mathbf{GCN}(V, W) = W_r(WVW_g) + V, \tag{7}$$

where W_g is the $D \times D$ dimension parameter matrix of a GCN layer, which contains inference weights between graph nodes. $W_r \in \mathbb{R}^{k \times k}$ is the weight matrix of the residual structure. The output of GCN, denoted by $V^* = \{v_1^*, v_2^*, \ldots, v_k^*\}, V_i^* \in \mathbb{R}^D$, is a new graph structure that contains inferential relationship between nodes. Applying such a GCN model to the visual graphs, we obtain:

$$\begin{aligned} V_{sp}^* &= \mathbf{GCN}_{sp}(V_{sp}, W_{sp}), \\ V_{se}^* &= \mathbf{GCN}_{se}(V_{se}, W_{se}). \end{aligned} \tag{8}$$

The semantic graph contains semantic relationships between regions, while the spatial graph emphasizes the inter-relation between the overlapped regions. The two graphs are combined together to represent the image. The final visual graph V_I^* is the average of the spatial graph V_{sp}^* and semantic graphs V_{se}^*:

$$V_I^* = \frac{V_{sp}^* + V_{se}^*}{2}. \tag{9}$$

As each graph node contributes to the global embedding differently, the graphs V_I^* is fed into a GRU to obtain final image embedding I, which is the last hidden state of GRUs:

$$I = GRU(V_I^*)_m. \tag{10}$$

Textual Reasoning and Embedding. The same to the image, GCN and GRU are employed for textual graph reasoning and embedding on V_t:

$$\begin{aligned} V_t^* &= \mathbf{GCN}_t(V_t, W_t), \\ T &= GRU(V_t^*)_n. \end{aligned} \tag{11}$$

The graph embedding I and the sentence embedding T are the final representations of the image and text, which are used for similarity computation during the test.

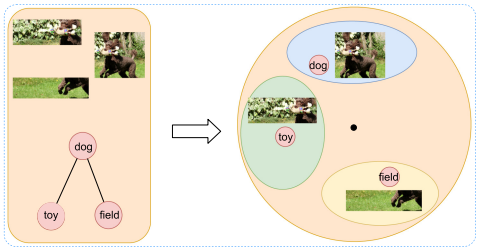

Fig. 4. Schematic diagram of graph node matching. The purpose of graph node matching loss is to achieve fine-grained matching between words in sentences and regions in pictures while training.

3.4 Loss Function

To better learn fine-grained cross-modal alignment and inter-relation reasoning without sacrificing the computational efficiency in the test, we use a combination of a graph node matching loss and an embedding loss during training.

Graph node matching loss. To learn fine-grained cross-modal correspondence and achieve inter-relation reasoning, a graph node matching loss is employed during training to promote region-word alignment, as shown in Figure 4. The node matching loss aims to make the distances between words and similar meanings regions in the image closer and makes the differences between different expressions larger. The loss is to calculate the sum of the maximum node similarity, which is defined as:

$$L_G(V_I^*, V_t^*) = \max[\triangle - m_{node}(V_I^*, V_t^*) + m_{node}(V_I^{*-}, V_t^*), 0], \tag{12}$$

where $\triangle$ is a margin parameter. Note that, since texts contain less information than images that have background information or indescribable objects, the nodes of the texture graph are less than those of the visual graph. Thus, only the text-to-image graph node matching is considered.

The graph node matching function $m_{node}(\cdot, \cdot)$ calculates the sum of the maximum similarity value in each node matching:

$$m_{node}(V_I, V_t) = \sum_{0 \le i \le m, 0 \le j \le l} \max[cos(V_{Ii}, V_{tj}), 0], \tag{13}$$

where $cos(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\alpha}^T \cdot \boldsymbol{\beta}}{|\boldsymbol{\alpha}||\boldsymbol{\beta}|}$ denotes the cosine function. Note that node-level ground-truth is not available in the graph node matching loss. The matching between nodes is learned potentially along with the image-sentence pair learning.

Graph embedding matching loss. Hard triplet loss [11] is used for embedding loss, which is to make the distance between the hard negative items and positive targets largest in the mini-batch. The embedding loss is defined as:

$$L_M(I, T) = \max[\gamma - cos(I, T) + cos(I^-, T), 0] + \max[\gamma - cos(I, T) + cos(I, T^-), 0], \tag{14}$$

where γ is a margin parameter. I^- and T^- are the hardest negatives to I and T in a mini-batch, respectively.

Overall loss. The final loss for training is a combination of the embedding loss and the graph node matching loss is:

$$L = L_M + \frac{L_G}{K}, \tag{15}$$

where K is a weighted hyperparameter, and $K = 3$ is used in our proposed model.

3.5 Comparison with Existing Graph-based Methods

Our CGMN model is closely related to several recent graph-based methods, such as the independent representation methods VSRN [24] and cross-interaction method GSMN [28].

In VSRN [24], the graph is used for visual intra-relation reasoning in image embedding extraction, but not used in sentences. Sentences are simply processed by GRUs and not represented as graphs to extract intra-sentence relationships. GSMN [28] represents both images and sentences by graphs, which focuses on node-level and structure-level correspondence for inter-relation alignment, but the used DNN-based similarity measure is computationally expensive.

Our CGMN is different from the above methods in graph construction and matching approach. Specifically, in CGMN, an image is simultaneously represented by a visual semantic graph and a spatial graph. Moreover, a graph node matching loss is proposed for fine-grained correspondence and inter-relation reasoning between images and sentences, which is only used during training.

4 EXPERIMENT

To evaluate the effectiveness of our proposed CGMN model, we test the image retrieval and sentence retrieval separately on the three most commonly used datasets, Flickr8K [55], Flickr30K [55], and MS-COCO [27]. Besides, our proposed CGMN model is compared with other SOTA models.

4.1 Datasets and Evaluation Metrics

4.1.1 Dataset. we test image retrieval and sentence retrieval separately on the three most commonly used datasets, Flickr8K [55], Flickr30K [55], and MS-COCO [27].

Flickr8K contains 8,000 images with five sentences for each image. Following References [4, 36], the dataset is split into three parts: 6,000 images for training, 1,000 images for validation, and 1,000 images for testing.

Flickr30K contains 31,783 images with five sentences for each image. Following Reference [11], we split the dataset into three parts: 29,783 images for training, 1,000 images for validation, and 1,000 images for testing.

MS-COCO consists of 113,287 images for training, 1,000 images for validation, and 5,000 images for testing. Similarly, each image has five annotations. The test results on MS-COCO are obtained by the average of five-fold cross-validation (named MS-COCO 1K) and full 5K test set (named MS-COCO 5K).

4.1.2 Evaluation Metrics. For both sentence retrieval and image retrieval, we evaluate the performance of the compared methods in terms of Recall at K (R@1, R@5, R@10), which defines the correct retrieval results are in the top-K ranking results. Meanwhile, we also use "Rsum" to reasonably represent the quality of a model, which is defined as:

$$Rsum = \underbrace{R@1 + R@5 + R@10}_{\texttt{Sentence retrieval}} + \underbrace{R@1 + R@5 + R@10}_{\texttt{Image retrieval}}. \tag{16}$$

When comparing with the current SOTA methods, we follow the same strategy commonly used in SOTA methods [4, 22, 24, 28] to combine results from two trained CGMN models by averaging their predicted similarity scores.

4.2 Implementation Details

The proposed CGMN is trained on the training set and validated on the validation set every 500 mini-batches and every epoch. For image representation, we use Faster R-CNN with bottom-up attention to describe the images, which consist of 36 regions with a 2,048-dimensional vector to describe each region for the top confidence. The IoU threshold ξ is set as 0.4. And for sentences, the

Table 1. Quantitative Results of Sentence Retrieval and Image Retrieval on Flickr30K

Type	Model	Sentence Retrieval			Image Retrieval			Rsum
		R@1	R@5	R@10	R@1	R@5	R@10	
Cross	SCAN [22]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
	RDAN [15]	68.1	91.0	95.9	54.1	80.9	87.2	477.2
	MMCA [50]	74.2	92.8	96.4	54.8	81.4	87.8	487.4
	CAAN [59]	70.1	91.6	97.2	52.8	79.0	87.9	478.6
	IMRAM [4]	74.1	93.0	96.6	53.9	79.4	87.2	484.2
	GSMN [28]	76.4	**94.3**	**97.3**	57.4	82.3	89.0	496.8
Independent	DVSA [18]	22.2	48.2	61.4	15.2	37.3	50.5	234.8
	m-RNN [35]	41.0	73.0	83.5	29.0	42.2	77.0	345.7
	DSPE [44]	50.1	79.7	89.2	39.6	75.2	86.9	420.7
	2WayNet [10]	49.8	67.5	-	36.0	55.6	-	208.9
	VSE++ [11]	52.9	-	87.2	39.6	-	79.5	259.2
	DPC [60]	55.6	81.9	89.5	39.1	69.2	80.9	416.2
	SCO [16]	55.5	82.0	89.3	41.1	70.5	80.1	418.5
	SGM [45]	71.8	91.7	95.5	53.5	79.6	86.5	478.6
	VSRN [24]	71.3	90.6	96.0	54.7	81.8	88.2	482.6
	CGMN (ours)	**77.9**	93.8	96.8	**59.9**	**85.1**	**90.6**	**504.1**

"Cross" denotes cross-interaction matching methods, while "Independent" denotes independent representation matching methods. The underlined numbers represent the best results in independent representation matching methods, and the bold numbers represent the best results in all methods.

words are embedded to 300-dimensional vectors and then encoded to 2,048-dimensional vectors. The number of GCN layers of the textual graph is 1 and that of the semantic graph and spatial graph are 4 and 2, respectively. The learning rate is 5e-4 and 2e-4 for Flickr30K and MS-COCO, respectively, in the first 10 epochs, and 90% off in every 5 epochs. The model is trained for totally 20 epochs. The batch size is 128 and the optimizer is Adam [37]. The margin parameter γ and $\triangle$ are both set as 0.2. The best model is selected through "Rsum" in the validation process. All our experiments are carried out on a single NVIDIA RTX 2080Ti GPU. The model is implemented by PyTorch.

4.3 Comparison with the State-of-the-Art Approaches

We compare our CGMN model with the current SOTA models on the three common benchmarks. In addition, we split the current apporaches into cross-combined matching methods (referred to as "Cross"), i.e., SCAN [22], RDAN [15], SGM [45], MMCA [50], CAAN [59], IMRAM [4], GSMN [28], and independent representation matching methods (referred to as "Independent"), i.e., DVSA [18], m-RNN [35], DSPE [44], 2WayNet [10], VSE++ [11], DPC [60], SCO [16], VSRN [24], and CGMN are compared with these two kinds of approaches, respectively. The results are shown in Tables 3, 1, and 2.

4.3.1 Result on Flickr30K. Table 1 presents the results on Flickr30K. The underlined number represents the best in independent representation matching methods, and the bold number represents the best in all methods. It can be observed that our model outperforms state-of-the-arts and achieves the best accuracy except for R@5 and R@10 in the sentence retrieval. Taking R@1 as an example, our model outperforms the current state-of-the-art model GSMN [28] by 1.5% in sentence retrieval and 2.5% in image retrieval, and the improvement in Rsum is 7.3. Moreover, our CGMN achieves the best in the independent representation matching methods. Compared with the cur-

Table 2. Quantitative Results of Sentence Retrieval and Image Retrieval on MS-COCO

Type	Model	Sentence Retrieval			Image Retrieval			Rsum
		R@1	R@5	R@10	R@1	@R5	R@10	
		MS-COCO 1K testset results						
Cross	SCAN [22]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
	RDAN [15]	74.6	**96.2**	**98.7**	61.6	89.2	94.7	515.0
	MMCA [50]	74.8	95.6	97.7	61.6	89.8	95.2	514.7
	CAAN [59]	75.5	95.4	98.5	61.3	89.7	95.2	515.6
	IMRAM [4]	76.7	95.6	98.5	61.7	89.1	95.0	516.6
	GSMN [28]	**78.4**	**96.4**	**98.6**	63.3	90.1	95.7	**522.5**
Independent	DVSA [18]	22.2	48.2	61.4	15.2	37.3	50.5	234.8
	m-RNN [35]	41.0	73.0	83.5	29.0	42.2	77.0	345.7
	DSPE [44]	40.3	68.9	79.9	29.7	60.1	72.1	351.0
	2WayNet [10]	49.8	67.5	-	36.0	55.6	-	208.9
	VSE++ [11]	64.7	-	95.9	52.0	-	92.0	304.6
	DPC [60]	65.6	89.8	95.5	47.1	79.9	90.0	467.9
	SCO [16]	69.9	92.9	97.5	56.7	87.5	94.8	499.3
	SGM [45]	73.4	93.6	97.8	57.5	87.3	94.3	504.1
	VSRN [24]	76.2	94.8	98.2	62.8	89.7	95.1	516.8
	CGMN (ours)	76.8	95.4	98.3	**63.8**	**90.7**	**95.7**	520.7
		MS-COCO 5K testset results						
Cross	SCAN [22]	50.4	82.2	90.0	38.6	69.3	80.4	410.9
	MMCA [50]	**54.0**	82.5	90.7	38.7	69.7	80.8	416.4
	CAAN [59]	52.5	**83.3**	90.9	**41.2**	70.3	**82.9**	**421.1**
	IMRAM [4]	53.7	83.2	**91.0**	39.7	69.1	79.8	416.5
Independent	DVSA [18]	16.5	39.2	52.0	10.7	29.6	42.2	190.2
	VSE++ [11]	41.3	-	81.2	30.3	-	72.4	225.2
	DPC [60]	41.2	70.5	81.1	25.3	53.4	66.4	337.9
	SCO [16]	42.8	72.3	83.0	33.1	62.9	75.5	369.6
	SGM [45]	50.0	79.3	83.9	35.3	64.9	76.5	393.9
	VSRN [24]	53.0	81.1	89.4	40.5	70.6	81.1	415.7
	CGMN (ours)	53.4	81.3	89.6	**41.2**	**71.9**	82.4	419.8

"Cross" denotes cross-interaction matching methods, while "Independent" denotes independent representation matching methods. The underlined numbers represent the best results in independent representation matching methods, and the bold numbers represent the best results in all methods.

rent independent representation matching method VSRN [24], our CGMN gains an improvement of 6.6% and 5.2% on R@1 in sentence retrieval and image retrieval, respectively. The results on R@5 and R@10 in the sentence retrieval task show no advantage, which may be due to that there exit more distracting targets or similar targets in the test set. Noteworthily, CGMN makes significant progress in sentence-to-image retrieval compared with state-of-the-arts. The reason will be analyzed later in the ablation study section.

4.3.2 Result on MS-COCO. Typically, there are two ways to conduct an image-sentence retrieval test on MS-COCO, including a five-fold cross-validation test (1K) and a full set test (5K). Table 2 shows the results on MS-COCO. Clearly, our model also performs well on the bigger dataset MS-COCO, which demonstrates the good generalization capability of our model. In MS-COCO 1K, it can be observed that our model outperforms current SOTA independent representation matching

Table 3. Quantitative Results of Sentence Retrieval and Image Retrieval on Flickr8K

Type	Model	Sentence Retrieval			Image Retrieval			Rsum
		R@1	R@5	R@10	R@1	R@5	R@10	
Cross	DeViSE [13]	4.8	16.5	27.3	5.9	20.1	29.6	104.2
	m-CNN [33]	24.8	53.7	67.1	20.3	47.6	61.7	275.2
	SCAN [22]	52.2	81.0	89.2	38.3	67.8	78.9	407.4
	IMRAM [4]	54.7	**84.2**	91.0	41.0	69.2	**79.9**	420.0
Independent	DVSA [18]	16.5	40.6	54.2	11.8	32.1	44.7	199.9
	VSRN* [24]	50.6	78.4	85.5	38.0	65.7	74.7	392.9
	CGMN (ours)	**57.0**	83.8	**91.4**	**41.3**	**69.6**	78.0	**421.2**

"Cross" denotes cross-interaction matching methods, while "Independent" denotes independent representation matching methods. The underlined numbers represent the best results in independent representation matching methods, and the bold numbers represent the best results in all methods. Note that, as the results of VSRN [24] are not reported on Flickr8K, we show the results by running codes provided by the authors.

methods and achieves competitive results to the best cross-interaction matching methods. Compared with SOTA independent representation matching method VSRN [24], our model attains 0.6% and 1.0% improvement on R@1 for MS-COCO in sentence retrieval and image retrieval, respectively. Similarly, the Rsum also increased by 3.9. Meanwhile, our CGMN is competitive to the best cross-interaction method GSMN [28] and outperforms other cross-interaction methods. Compared with GSMN [28], our CGMN achieves better accuracy on image retrieval and performs not as well as it. Moreover, in the bigger MS-COCO 5K test set, CGMN yields the best results in image retrieval and sentence retrieval in all independent representation matching methods and equally good results with cross-interaction methods. On R@1, Our CGMN outperforms VSRN [24] for 0.4% and 0.7% in sentence retrieval and image retrieval, respectively.

4.3.3 Result on Flickr8K. Table 3 presents the results on Flickr8K. Note that Flickr8K is an antiquated dataset that has been enlarged to Flickr30K, so most recent works are not validated on Flickr8K. Nevertheless, Flickr8K can still be used to validate the effectiveness of a network in the condition of a small dataset. It can be observed that our model outperforms current methods. Since Flickr8K is a small dataset with only 6,000 images for training, independent representation matching methods regularly perform worse than cross-interaction matching methods. For example, VSRN [24] performs much better than SCAN [22] in both Flickr30K and MS-COCO but gains a 14.5% decrease on Rsum in Flickr8K. However, our CGMN outperforms IMRAM [4] for 1.2% increase on Rsum. Especially, our CGMN gains 2.3% and 0.3% on R@1 in sentence retrieval and image retrieval, respectively.

4.3.4 Analysis of the Results. With the test results, we can analyze the results and get some findings as follows:

(a) In general, cross-interaction matching methods perform better than independent representation matching methods on both sentence retrieval and image retrieval. In the cross-interaction methods, the interaction between images and text in extracting embedding helps to better learn potential relevance, hence helps fine-grained matching. Our CGMN uses a novel graph node matching loss to learn a fine-grained region-word correspondence and inter-relation reasoning during training, which achieves comparable good or better matching results as cross-interaction matching methods. The results show that inter-relation reasoning plays an important role in the matching of images and sentences.

(b) Compared with the previous independent representation matching methods, our CGMN also uses embedding matching in test. It shows that adding fine-grained correspondence in training can help the final embedding generation. Different from the generative model in VSRN [24] that benefits accurate matching, our CGMN uses a novel graph node matching loss to achieve correspondence and inter-relation reasoning between image regions and words in sentences, which benefits final embedding matching during retrieval.

4.4 Analysis on Computational Complexity

In addition to retrieval accuracy, retrieval efficiency is another important aspect of performance, especially in practical applications. We analyze the computational complexity of our CGMN to show the superiority in retrieval efficiency over other cross-interaction matching methods. Moreover, we conduct experiments on offline retrieval and online retrieval to demonstrate its efficiency. Finally, the complexity of the training process is analyzed.

In offline retrieval, the retrieval process only includes query processing and similarity calculation, and the efficiency mainly depends on the matching method. The computational complexity of offline retrieval is $O(K \times n)$, where K depends on the complexity of a single matching between a query and an item in the gallery, and n is the number of items in the gallery. In the independent representation matching methods, the items in the gallery have been already processed to embeddings, which are stored in the database. Then, the cosine similarity is used as the measurement of the similarity between the embedding of the query and those of all items in the gallery. In this kind of method, $K = O(D)$, where D is the dimension of the embeddings. However, in other cross-interaction methods, the embeddings can not be pre-stored. Therefore, for a query, all the items in the gallery need to be processed along with the query in a network-based method, which leads to high computational complexity. In such methods, $K = O(MD)$, where M depends on the complexity (Flops) of the network for matching and $M \gg 2$. Note that the complexity of two kinds of cross-combined matching methods (Figure 1(b1) and (b2)) is the same. Compared with methods without cross-embedding extraction of the same complexity M, the methods with cross-embedding extraction take one more step in embedding matching; the complexity is $K = O((1 + M)D) = O(MD)$. Taking SCAN as an example: Each image is represented as an $m \times D$-dimensional embedding, while a sentence is represented as an $l \times D$-dimensional embedding, where m and l denote the number of regions and the length of a sentence. The complexity of matching between images and sentences is $K = O(ml \times D)$, where $M = ml$ is the complexity of the network. Therefore, our CGMN is more efficient than the cross-interaction methods in offline retrieval.

In online retrieval, the retrieval process also includes the processing of data in the gallery. The computational complexity of online retrieval is $O(K \times n + G \times n)$, where K and n are the same as those in offline retrieval, and G depends on the computational complexity of embedding generating networks. Take a test set with n queries and m items in the gallery as an example, in cross-interaction matching methods, the complexity is $O(MD \times nm + G_c \times (n + m))$. However, in independent representation methods, the complexity is $O(D \times nm + G_i \times (n + m))$, which is much less than cross-interaction matching methods. Therefore, there is also an advantage in the efficiency of our CGMN over the cross-interaction methods in online retrieval.

Experiments of bi-directional cross-modal retrieval are conducted on the Flickr30K test set to compare the retrieval efficiency in offline retrieval and online retrieval between our CGMN and other cross-interaction matching methods, i.e., SCAN [22], IMRAM [4], GSMN [28]. The Flickr30K test set consists of 1,000 images and 5,000 sentences. The experiments are carried out on an Intel I9-9900KF CPU and a single NVIDIA RTX 2080Ti GPU. It can be observed in Table 4 that our CGMN spends only 3.6 seconds in offline retrieval and 10.6 seconds in online retrieval, while the retrieval runtime of the other models is much greater than ours. Compared with the most

Table 4. Runtime Comparison for a Complete Retrieval on Flickr30K Test Set

Type	Model	Offline retrieval time (s)	Online retrieval time (s)
Cross	SCAN [22]	485.6	497.1
	IMRAM [4]	1,112.0	1,117.0
	GSMN [28]	222.2	229.4
Independent	VSRN [24]	3.5	9.2
	CGMN (ours)	3.5	10.6

Table 5. Quantitative Results of the Time Cost of Each Parts

Graph-based Network	Graph Node Matching Loss	Average Train Time (s/Batch)
×	×	0.17
✓	×	0.55
×	✓	0.42
✓	✓	0.80

efficient cross-interaction method GSMN in the experiments, our CGMN achieves about 61.7 times and 21.6 times faster in offline and online bi-directional image-sentence retrieval, respectively. In addition, compared with an independent representation matching method VSRN [24], our CGMN spends the same time in offline retrieval and achieves only 1.15 times slower in online retrieval. The results demonstrate that the proposed CGMN has a remarkable advantage in both offline and online retrieval efficiency over other cross-interaction matching methods and is as efficient as other independent representation matching methods.

In terms of the training process, the complexity is similar to that of the retrieval process. As for region detection, the Faster R-CNN is pre-trained in the Visual Genome [21] and is used to extract region-level features for every image before training. That is, Faster R-CNN does not participate in the training process and affects training efficiency. Although graph-based networks are of higher complexity than common CNN-based networks, the efficiency is not affected much. Experiments are conducted to test the time cost of each part. Table 5 shows that a network with a graph-based reasoning module and graph node matching loss cost five times the time for training in a mini-batch. Considering the improvement in effectiveness, the efficiency of training is acceptable.

4.5 Ablation Study

4.5.1 Analysis of Each Part in CGMN. We conduct an ablation study on Flickr30K to verify the effectiveness of each part in CGMN. These parts include the graph node matching module, the spatial graph, the sentence graph, and the semantic graph, which are referred to as "matching," "P," "S," and "R," respectively. Note that the graph node matching module does not apply in the two cases of "w/o S," for there is no graph constructed in sentence processing. Table 6 shows the R@1, R@5, and R@10 of image retrieval and sentence retrieval, and Rsum results on Flickr30K. It is clear that when any part is removed, the matching performance degrades.

(a) The impact of the spatial graph and the semantic graph is different, and both of them are beneficial to the image embedding for retrieval, while the combination of them yields better performance. Spatial graphs and semantic graphs focus on different aspects, and they can focus on the relationship between semantics and the relationship between relative positions in the images, indicating that both semantic connections and positional relationships play an important role in image embedding and retrieval.

Table 6. Ablation Studies of Our Model for Bidirectional Image-sentence Retrieval on Flickr30K

Model	Sentence Retrieval			Image Retrieval			Rsum
	R@1	R@5	R10	R@1	R@5	R10	
w/o S & R & P	71.3	91.0	95.4	57.5	83.1	89.6	487.9
w/o R & P	74.8	91.6	95.6	58.0	84.0	90.3	494.3
w/o S & R	74.2	92.0	95.3	57.2	83.5	89.9	492.1
w/o S & P	71.7	93.4	95.6	57.6	83.6	90.2	492.1
w/o S	72.5	92.9	95.9	57.5	83.5	90.4	492.7
w/o R	74.0	**94.1**	96.5	59.0	85.0	90.5	499.1
w/o P	77.0	93.5	96.6	**60.1**	**85.1**	90.5	502.8
w/o matching	73.3	92.6	96.0	59.3	84.6	89.7	495.5
Full model	**77.9**	93.8	**96.8**	59.9	**85.1**	**90.6**	**504.1**

Table 7. Quantitative Results of Our Method with Different Embedding Functions of Images and Sentences

Number	Embedding Function		Sentence Retrieval			Image Retrieval			Rsum
	image	sentence	R@1	R@5	R@10	R@1	R@5	R@10	
1	mean pooling	mean pooling	69.8	90.8	95.4	55.2	82.8	89.0	483.0
2	mean pooling	GRU	74.3	93.0	96.0	**60.6**	84.6	**90.6**	499.1
3	GRU	mean pooling	74.2	92.4	96.4	58.9	83.9	90.2	496.1
4	GRU	GRU	**77.9**	**93.8**	**96.8**	59.9	**85.1**	**90.6**	**504.1**

(b) The results show that the textual graph distinctly improves retrieval performance, particularly in image retrieval. Sentence graph is built via semantic dependency and semantic similarity, which express the connection between words from the structure of natural language. When applying GCNs on the sentence graph, it benefits the inter-relation reasoning of words. Moreover, it emphasizes the relationship between different entities, including actions and locations, which benefits the sentence embedding.

(c) The graph node matching loss also contributes significantly to improve the performance. Without the graph node matching loss, the R@1 of sentence retrieval and image retrieval decline 4.6% and 0.6%, respectively. The graph node matching loss helps to better learn the fine-grained correspondence between regions of images and words in sentences, which benefits the final embedding matching. It indicates that our novel graph node matching loss is effective for cross-modal retrieval.

4.5.2 Analysis of the Embedding Methods. We also test the results of the embedding method of graph representation. The embedding methods are mean pooling for the average of graph nodes, while the other used in our CGMN is GRU. Results shown in Table 7 indicate that the effect of mean pooling is not as good as GRU, while GRUs achieve better results in both two graphs. We believe that it is because different nodes of the graphs contribute differently to the final embedding. GRUs benefit from intra-relation reasoning between these nodes to achieve more accurate matching.

4.5.3 Analysis of the Effectiveness of the Pair-wise Similarity and Semantic Dependency in Graph Construction. We also test the effectiveness of the similarity measure in the spatial graph and textual graph, which is mentioned in Equations (4) and (6). We test the region similarity in the spatial graph, as well as semantic dependency and word similarity in the textual graph. Note that we do not test the semantic graph, as its effect has been evaluated above, and the effect of IoU is

Table 8. Quantitative Results of Our Method on the Effectiveness of the Node Similarity

Spatial Graph		Textual Graph		Sentence Retrieval			Image Retrieval			Rsum
IoU	Similarity	Dependency	Similarity	R@1	R@5	R@10	R@1	R@5	R@10	
✓		✓		75.1	93.6	96.1	59.9	84.2	90.4	499.3
✓			✓	75.1	93.4	96.7	60.0	84.8	90.3	500.3
✓		✓	✓	74.3	93.1	95.9	60.6	84.6	90.6	499.1
✓	✓	✓		75.2	**94.3**	96.4	**60.4**	84,7	90.5	501.5
✓	✓		✓	74.1	92.9	**96.8**	59.9	84.9	**90.7**	499.3
✓	✓	✓	✓	**77.9**	93.8	**96.8**	59.9	**85.1**	90.6	**504.1**

Table 9. Quantitative Results of Our Method on the Effectiveness of the Combination Methods

Number	Combination Methods	Sentence Retrieval			Image Retrieval			Rsum
		R@1	R@5	R@10	R@1	R@5	R@10	
1	gated combination	76.0	92.5	96.3	58.9	84.1	90.2	498.0
2	weight combination	75.2	93.8	97.0	59.3	85.2	91.0	501.5
3	mean pooling	**77.9**	**93.8**	**96.8**	**59.9**	**85.1**	**90.6**	**504.1**

Table 10. Quantitative Results of Our Method for Different Weighted Hyperparameter K

$K =$	Sentence Retrieval			Image Retrieval			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
1	75.3	93.6	96.5	59.9	85.0	90.3	500.7
2	76.8	94.0	96.8	60.2	85.3	90.8	503.9
3	77.9	93.8	96.8	59.9	85.1	90.6	504.1
4	77.8	93.3	96.4	59.7	84.9	90.3	502.4

not tested, since region similarity can not represent the spatial relation, where the spatial graph turns to be the same as a semantic graph without IoU. The results in Table 8 show the effectiveness of each part, and the combination of these parts achieves the best retrieval results.

4.5.4 Analysis of the Combination Methods of the Spatial Graph and Semantic Graph. For a better combination of spatial graph and semantic graph in image embedding, we compare multiple methods, which are gated combination (the two graphs are combined by a gate union), weight combination (the weight of edges are combined before graph processed by GCNs), and mean pooling used in our CGMN, which are introduced in detail in Appendix A. Table 9 shows the results, where the mean pooling method achieves the best accuracy.

4.5.5 Analysis of Loss Ratio for Training. We further evaluate the effect of the loss balance parameter K. Since matching during retrieval is achieved through embedding matching, the graph node matching loss is an auxiliary task to learn fine-grained correspondence and inter-relation reasoning. We test the parameter K by different values, e.g., $K \in \{1, 2, 3, 4\}$. Table 10 presents the quantitative results. It can be observed that $K = 3$ yields the best performance.

4.6 Visualization Analysis

We visualize the graph node matching results to test whether our graph matching loss helps to learn fine-grained image-text matching. Figure 5 provides a visualization of the region-word correspondence results. The plotted bounding boxes are the top three ranked regions similar to

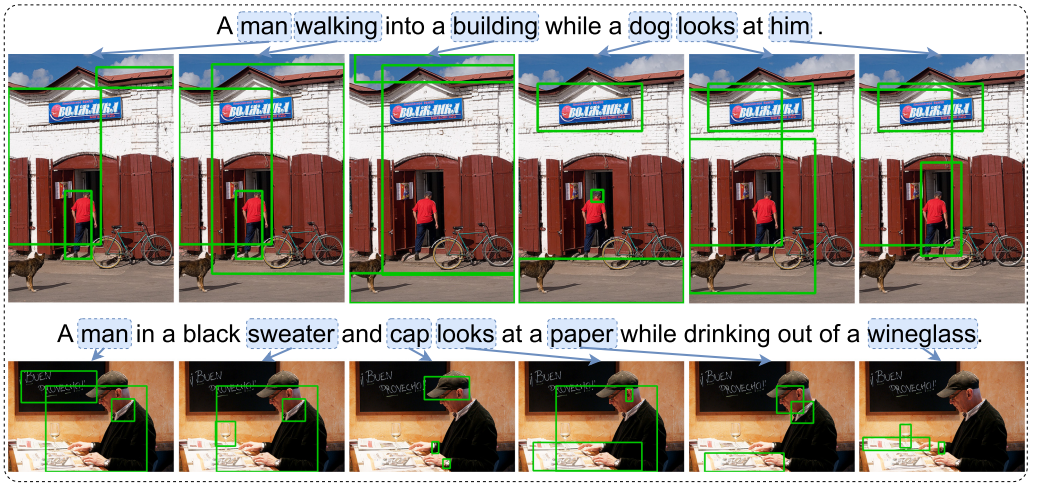

Fig. 5. Result visualization on region-word pair correspondence. The words in the blue box correspond to the regions in the image, and the top three regions closest to the words are selected as fine-grained matching according to formula 14.

the target words. As for the example above, the items "man," "building," and "dog" correspond to the regions in the image well. Besides, the verbs "walking" and "looks" refer to some regions that express the action logically. In particular, the word "him" correctly points to the region of the human. Moreover, in the matching query below, all items also match the regions in the image well.

For further validating the effect of graph node matching loss, Figure 6 provides a comparison between CGMN with and without graph node matching. In Figure 6, the retrieval results are displayed, where the numbers represent the order of results. It can be observed that the network with graph node matching loss achieves a better performance with fine-grained alignment. Take the first sentence for example. All the images partially correspond to the sentences, e.g., red jacket, jeans, or bench. However, these objects do not appear in other images together and glasses only exist in the right images. By graph node matching, networks reduce the neglect of multiple objects. In summary, the region-word correspondence performs well. Our novel graph node matching loss helps to learn fine-grained cross-modal correspondence, which is helpful for embedding matching.

Figures 7 and 8 further provide visualization of text-to-image retrieval and image-to-text retrieval on Flickr30K. In Figure 7, the top-ranked images are displayed, where the matched images are in green boxes, while mismatched are in red boxes. The top one ranked results are correctly matched, and the rest are also very similar to the semantic content of the texts. Taking the first sentence as an example, the model even matches the image regions with the word "carpet," instead of just matching two dogs. In Figure 8, the top five most similar sentences are shown in the list, where correct results are in green, while wrong ones are in red. It can be observed that our CGMN performs well in sentence retrieval.

5 CONCLUSION AND FUTURE WORK

We propose a **Cross-modal Graph Matching Network (CGMN)** for image-text retrieval, which constructs graphs for both image and text. The visual graph is constructed to discover the visual and spatial connection, while the textual graph is constructed to extract semantic connection.

Fig. 6. The retrieval results of CGMN with or without graph node matching. The top four most similar images are shown, while the numbers represent the order of results, and the true result is marked with green. In addition, the correspondence between regions and words is in the same colors.

Fig. 7. Results of image retrieval on Flickr30K. We show the top-ranked images in the dataset, where the matched images are in green boxes, while the mismatched ones are in red boxes.

Further, we use a graph node matching loss during training to learn fine-grained cross-modal correspondence. Experiments on three common benchmarks, Flickr8K, Flickr30K, and MS-COCO, show that CGMN outperforms SOTA independent methods and achieves competitive results to cross-interaction methods. Particularly, CGMN is far more efficient than the current SOTA methods involving interactively similarity measurement. Further, we do some ablation studies to prove the effectiveness of each part in our CGMN model. Visualization of the results shows the fine-grained cross-modal correspondence can be achieved by the graph node matching. In the future, we will explore the methods of inter-relation extraction and reasoning in image-text matching.

Fig. 8. Results of sentence retrieval on Flickr30K. The top five most similar sentences are shown in the list, where correct results are in green while wrong ones are in red.

APPENDIX

A THE FORMULAS IN ABLATION STUDY

Here is the detailed introduction of equations in the ablation study of "Analysis of the combination methods of the spatial graph and semantic graph." The combination methods are divided as gated combination, weight combination, and mean pooling. The gated combine is defined as:

$$\begin{aligned} V_1 &= W_1 V_{se}^*, V_2 = W_2 V_{sp}^*, \\ t &= \sigma(U_1 V_1 + U_2 V_2), \\ V_I^* &= t \cdot V_1 + (1 - t) \cdot V_2, \end{aligned} \tag{17}$$

where W and U are the parameters of a fully connected layer, σ is a sigmoid function.

Different from other methods that combine the graphs after graph reasoning, the weight is combined before graph reasoning, which is defined as:

$$W = W_{se} + W_{sp}. \tag{18}$$

Then, the combined visual graph is fed to GCN for relation reasoning.

The mean pooling method is defined in Equation (9).

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 (CVPR)*.

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2017. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Yunsheng Bai, Hao Ding, Song Bian, Ting Chen, Yizhou Sun, and Wei Wang. 2019. SimGNN: A neural network approach to fast graph similarity computation. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. 384–392.

[4] 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 (CVPR)*.

[5] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020. Graph optimal transport for cross-domain alignment. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1542–1553.

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

[7] 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*. 10583–10590.

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *arXiv preprint arXiv:1412.3555*.
[9] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. *arXiv preprint arXiv:2101.01368*.
[10] Aviv Eisenschtat and Lior Wolf. 2017. Linking image and text with 2-way nets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[11] 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] Matthias Fey, Jan E. Lenssen, Christopher Morris, Jonathan Masci, and Nils M. Kriege. 2020. Deep graph matching consensus. In *Proceedings of the International Conference on Learning Representations*.
[13] Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marcaurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 2121–2129.
[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[15] Zhibin Hu, Yongsheng Luo, Jiong Lin, Yan Yan, and Jian Chen. 2019. Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*.
[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 (CVPR)*.
[17] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. 2019. Saliency-guided attention network for image-sentence matching. In *Proceedings of the International Conference on Computer Vision*.
[18] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3128–3137.
[19] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*.
[20] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations*.
[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Retrieved from https://arxiv.org/abs/1602.07332.
[22] 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*.
[23] Kuang-Huei Lee, Hamid Palangi, Xi Chen, Houdong Hu, and Jianfeng Gao. 2019. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators. *arXiv preprint arXiv:1909.09953*.
[24] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the International Conference on Computer Vision*.
[25] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision*. Springer, 121–137.
[26] Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, and Pushmeet Kohli. 2019. Graph matching networks for learning the similarity of graph structured objects. In *Proceedings of the International Conference on Machine Learning*. 3835–3845.
[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*.
[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 (CVPR)*.
[29] Fangyu Liu, Rémi Lebret, Didier Orel, Philippe Sordet, and Karl Aberer. 2020. Upgrading the newsroom: An automated image selection system for news articles. *ACM Trans. Multim. Comput., Commun. Applic* 16, 3 (2020), 1–28.
[30] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[31] Xiaoxiao Liu and Qingyang Xu. 2021. Adaptive attention-based high-level semantic introduction for image caption. *ACM Trans. Multim. Comput. Commun. Appl.* 16, 4 (2021), 128:1–128:22.

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 13–23.

[33] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision*. 2623–2631.

[34] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 55–60.

[35] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (M-RNN). In *Proceedings of the International Conference on Learning Representations*.

[36] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2017. Hierarchical multimodal LSTM for dense visual-semantic embedding. In *Proceedings of the IEEE International Conference on Computer Vision*. 1881–1889.

[37] Kingma Diederik P. and Ba Jimmy. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

[38] L. Peng, Y. Yang, Z. Wang, Z. Huang, and H. T. Shen. 2020. MRA-Net: Improving VQA via multi-modal relation attention network. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1 (2020), 1–1. DOI: https://doi.org/10.1109/TPAMI.2020.3004830

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

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*.

[41] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Trans. Assoc. Computat. Ling.* 2 (2014), 207–218.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

[43] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

[44] 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 (CVPR)*. 5005–5013.

[45] 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.

[46] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuanfang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[47] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*.

[48] 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/CVF International Conference on Computer Vision*. 5764–5773.

[49] Jonatas Wehrmann, Camila Kolling, and Rodrigo C. Barros. 2020. Adaptive cross-modal embeddings for image-text alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 12313–12320.

[50] 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 (CVPR)*.

[51] K. Wen, X. Gu, and Q. Cheng. 2020. Learning dual semantic relations with graph attention for image-text matching. *IEEE Trans. Circ. Syst. Vid. Technol.* 31, 7 (2020), 1–1. DOI: https://doi.org/10.1109/TCSVT.2020.3030656

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

[53] Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. 2019. Cross-lingual knowledge graph alignment via graph matching neural network. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*. 3156–3161.

[54] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[55] 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. *Trans. Assoc. Comput. Linguistics*. 2 (2014), 67–78.

[56] Weijie Yu, Chen Xu, Jun Xu, Liang Pang, Xiaopeng Gao, Xiaozhao Wang, and Ji-Rong Wen. 2020. Wasserstein distance regularized sequence representation for text matching in asymmetrical domains. *arXiv preprint arXiv:2010.07717*.
[57] Jin Yuan, Lei Zhang, Songrui Guo, Yi Xiao, and Zhiyong Li. 2020. Image captioning with a joint attention mechanism by visual concept samples. *ACM Trans. Multim. Comput. Commun. Appl.* 16, 3 (2020), 83:1–83:22.
[58] Andrei Zanfir and Cristian Sminchisescu. 2018. Deep learning of graph matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2684–2693.
[59] 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 (CVPR)*.
[60] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Trans. Multim. Comput. Commun. Appl.* 16, 2 (2020), 51:1–51:23.

Received May 2021; revised September 2021; accepted November 2021