2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Coarse-to-Fine Scene Graph Similarity Reasoning for Image-text Retrieval

Chengsong Sun[1], Qingyun Liu[2], Yuankun Liu[1], Boyan Liu[3], Xiang Yuan[1],

Bingce Wang[1], Tong Mo[1](✉), and Weiping Li [1]

[1]School of Software and Microelectronics, Peking University, Beijing 100871, China
`438700445@qq.com {liuyk, xiangyuan, wangbingce}@stu.pku.edu.cn,`
`{motong,wpli}@ss.pku.edu.cn`
[2] School of Computer Science and Technology, Beijing Jiaotong University, Beijing
100044, China `21281010@bjtu.edu.cn`
[3] College of Economics and Management,China Agricultural University, Beijing
100083, China `1635027294@qq.com`

Abstract. Image-text retrieval is a crucial task, which targets at finding the counterparts from the opposing modalities. Scene graph based image-text retrieval methods leverage the object and predicate features to reason the cross-modal similarity, therefore increasing the retrieval accuracy. However, existing scene graph based image-text retrieval methods simply fuse the similarity calculations for features at each granularity in a single network, which only brings a slight improvement in the retrieval performance. The features of the scene graph fail to be effectively utilized. Therefore, this paper proposes a **C**oarse-to-**F**ine **S**cene **G**raph Similarity **R**easoning (CFSGR) method to conduct coarse-grained and fine-grained cross-modal similarity reasoning, separately. CFSGR includes two networks: coarse-grained similarity reasoning network for graphs, fine-grained similarity reasoning network for objects and predicates. Moreover, CFSGR conducts local and global alignments for each feature, ensuring that the similarities at each granularity of visual and textual scene graphs are fully exploited. The evaluation and ablation study on Flickr30K demonstrates the superiority of CFSGR among the SOTA(State-Of-The-Art) image-text retrieval methods, and CFSGR achieves competitive results with $Rsum$ as 506. The source code is available at `https://github.com/okeike/CFSGR`.

Keywords: Image-text retrieval, Multi-modal similarity reasoning, Scene graph, Contrastive learning.

1 Introduction

Image-text retrieval, a cross-modal retrieval task, aims at bridging the gap between two heterogeneous modalities, images and texts, by measuring their cross-modal similarity. This task involves retrieving the most relevant image given a text query, or vice versa,

✉ Corresponding author

which requires understanding the complex relationship between visual and textual elements. Image-text retrieval plays a vital role in various computer vision and natural language processing tasks, such as image captioning [2, 18] and visual question answering [2, 16].

Scene graph based image-text retrieval methods are significant as they support both coarse-grained local and global similarities reasoning, as well as fine-grained object and predicate similarities reasoning, demonstrating crucial implications in image-text retrieval. For example, methods like SGM [15] focus on local matching of objects and their relationships, while LGSGM [11] additionally introduces global matching, aiming to align visual and textual graph structures. However, existing scene graph based methods do not effectively make use of the coarse-grained and fine-grained features to reason the cross-modal similarity, as they simply combine the similarity calculations for features at each granularity in a single network. This implementation causes the network easily converge to a local optimum, thus the features of scene graph are not fully leveraged.

To address this limitation in image-text retrieval, this paper introduces a novel Coarse-to-Fine Scene Graph Similarity Reasoning (CFSGR) method. The CFSGR framework is designed to enhance retrieval performance by integrating different local and global cross-modal alignments in coarse-grained and fine-grained similarity reasoning networks. Specifically, it designs coarse-grained graph similarity reasoning alongside fine-grained object and predicate similarity reasoning, ensuring a comprehensive understanding of scene graphs.

For local alignment in the coarse-grained similarity reasoning for graphs, CFSGR calculates cross-modal similarities for local features within scene graphs. In terms of coarse-grained global alignment, the CFSGR derives global features from local features and conducts comprehensive graph similarity reasoning, incorporating both global and local features. For fine-grained global alignment of objects and predicates, CFSGR calculates cosine similarities of their respective global representations. To achieve fine-grained local alignment, CFSGR performs detailed similarity reasoning based on the local and global representations of objects and predicates, individually.

The main contributions of this paper:

(1) This paper proposes a novel scene graph based image-text retrieval method CFSGR, which contains two networks: coarse-grained graph similarity reasoning network, fine-grained object and predicate similarity reasoning network. These two networks enable CFSGR to fully utilize the features at each granularity in scene graphs.

(2) CFSGR conducts local and global alignments in both the coarse-grained and fine-grained similarity reasoning networks, and the implementations of these alignments differ across the two granularities. For instance, at the coarse-grained level, local alignment is achieved through dot-product calculations. At the fine-grained level, the local alignment is realized using a detailed similarity reasoning module. By integrating these diverse alignment strategies, CFSGR achieves better performance.

(3) The proposed CFSGR achieves competitive results on a benchmark dataset Flickr30K [19], and it scores 506 in terms of $Rsum$, demonstrating superiority among scene graph based methods. The ablation study for CFSGR on Flickr30K [19] proves the effectiveness of the main components designed.

2 Related Work

Existing image-text retrieval methods can be divided into two main categories: object feature based methods and scene graph based methods. Object feature based methods extract object features from images, offering flexibility and computational efficiency in calculating the text-to-object similarity. Scene graph based methods typically construct a graph structure for both the image and text and then align the corresponding components (objects, predicates) across modalities. With these features, scene graph based methods can perform sophisticated reasoning.

2.1 Object Feature based Methods

Object feature based methods mainly focus on obtaining the local object features and global features. Unlike scene graph based methods, which explicitly model objects and relations as nodes to form a heterogeneous graph, these methods primarily rely on direct extraction of object features, and conduct similarity reasoning on them.

Object features based methods focus on matching fine-grained details between images and texts, such as individual objects and relations. Ji et al. [8] propose SHAN, a step-wise hierarchical alignment network that decomposes image-text retrieval into multi-step cross-modal reasoning process. Chen et al. [3] propose $VSE\infty$, a Generalized Pooling Operator (GPO) that autonomously adapts to the optimal pooling strategy for various features. Shi et al. [13] propose DCPA, a novel decoupled manner for training and inferencing, where image-to-sentence matching is executed in textual semantic space and sentence-to image matching is executed in visual semantic space. Wang et al. [17] propose WCGL to construct and encode graphs for cross-modal samples and utilize a Wasserstein coupled dictionary for feature transformation. Diao et al. [5] propose SGRAF to enhance image-text retrieval by learning vector-based similarity representations and using a graph-based reasoning module to infer relation-aware similarities. Cheng et al. [4] propose CGMN, leveraging fully-connected graphs for intra-relation reasoning and a novel graph node matching loss for inter-relation reasoning, which explores both intra- and inter-relations without introducing network interaction.

In summary, while object feature based methods achieve high efficiency in matching visual and textual representations, they often fall short in modeling the deep, relational semantics between objects and their interactions. Scene graph based methods, on the other hand, excel at capturing these relationships but can still benefit from more detailed and multi-level semantic reasoning, as demonstrated by our model, CFSGR.

2.2 Scene Graph based Methods

Scene graph based methods have gained significant attention in image-text retrieval due to their ability to capture detailed relationships between objects in both images and texts. These methods construct scene graphs for both modalities, explicitly modeling objects, their relationships, and attributes.

For instance, Wang et al. [15] propose SGM, focusing on local matching between objects and their relations within scene graphs. Nguyen et al. [11] propose LGSGM, which expands upon SGM [15] by incorporating global matching. Duan et al. [6] propose HSGMP, which introduces metapaths along with Heterogeneous Message Passing to extract semantic relationships and enhance cross-modal similarity measurement. Fan et al. . [7] propose SSAMT to construct scene graphs solely for textual data, integrating word, phrase, and sentence-level features. However, their model lacks a corresponding visual scene graph and does not perform reasoning based on scene graph structures. Pei et al. [12] propose SGSIN to conduct semantic inference within each modality independently, but their model fails to capture the multi-level hierarchical structure of scene graphs.

Above all, These scene graph based methods suffer from the limited utilization of scene graph features, which leads to suboptimal performance. To address this limitation, this paper proposes CFSGR, which is a hierarchical similarity reasoning framework that separately conducts coarse-grained and fine-grained cross-modal similarity reasoning, ensuring a more comprehensive and effective utilization of scene graph features.

3 Method

As is illustrated in Figure 1, CFSGR firstly builds the visual and textual scene graphs by using the off-the-shelf Neural Motifs [20] and SPICE [1] separately. Then, the object and predicate nodes in visual and textual scene graphs are fed into their corresponding embedding layers.

For visual scene graph $SG_I = (O_I, R_I)$, $O_I = o_{I1}, o_{I2}, \dots, o_{Iv_o}$ and $R_I = r_{I1}, r_{I2}, \dots, r_{Iv_r}$ respectively refer to the object and predicate sets. As shown in the visual scene graph of figure 1, the blue bounding boxes with their corresponding linguistic labels such as ``Woman" and ``Knife" denote the object nodes, while the linguistic labels on green arrows such as ``Hold" and ``On" denote the predicates between the source objects and target objects.

Upon acquiring the visual scene graphs, the object representations are obtained through fusing the features of bounding boxes $emb_{ob}(o_{Ii}) = W_{ob} h_{o_{Ii}}$and their corresponding linguistc labels $emb_{ol}(o_{Ii}) = W_{ol} h_{o_{Ii}}$, while the predicate representations are obtained through fusing the features of bounding-box unions of source object and target object $emb_{ru}(r_{Ik}) = W_{ru} h_{r_{Ik}}$ with linguistc labels of predicates $emb_{rl}(r_{Ik}) = W_{rl} h_{r_{Ik}}$:

$$\begin{aligned} emb(o_{Ii}) &= W_{IO}\left(concat\left(emb_{ob}(o_{Ii}), emb_{ol}(o_{Ii})\right)\right) \\ emb(r_{Ik}) &= W_{IR}\left(concat\left(emb_{ru}(r_{Ik}), emb_{rl}(r_{Ik})\right)\right) \end{aligned} \tag{1}$$

Textual scene graph $SG_T = (O_T, R_T)$ uses the $O_T = \{o_{T1}, o_{T2}, \dots, o_{Tt_o}\}$ and $R_T = \{r_{T1}, r_{T2}, \dots, r_{Tt_r}\}$ to separately refer to the object and predicate sets. As shown in the textual scene graph of Figure 1, the blue linguistic labels in rectangles such as ``Kitchen" and ``Cup" denote the object nodes, while the linguistic labels on green arrows such as ``Cut" and ``Surround by" denote the predicates between the source objects and target objects.

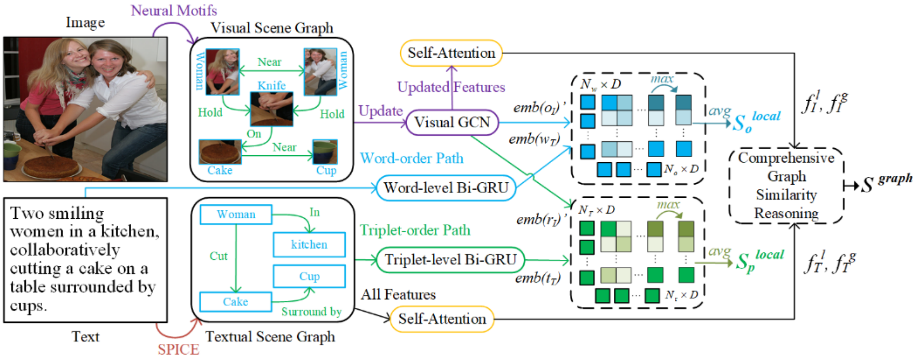

Fig. 1. Coarse-grained graph similarity reasoning of CFSGR. After building the visual and textual scene graphs, CFSGR calculates the local graph similarity reasoning based on the visual updated features $emb(o_I)', emb(r_I)'$ and textual local features $emb(w_T), emb(t_T)$. Then, CFSGR obtains visual global feature f_I^g and textual global feature f_T^g. Finally, CFSGR conducts comprehensive graph similarity reasoning with visual features f_I^l, f_I^g and textual featuresf_T^l, f_T^g.

After obtaining the textual scene graph, CFSGR encodes the word-order paths $emb(w_i)$ of input text and triplet-order paths $emb(t_i)$ of textual scene graph with their respective Bi-GRUs:

$$\begin{aligned} emb(w_{Ti}) &= \frac{\overrightarrow{GRU_w\big(emb(w_i)\big)} + \overleftarrow{GRU_w\big(emb(w_i)\big)}}{2} \\ emb(t_{Tk}) &= \frac{\overrightarrow{GRU_t\big(emb(t_k)\big)} + \overleftarrow{GRU_t\big(emb(t_k)\big)}}{2} \end{aligned} \tag{2}$$

3.1 Coarse-grained Graph Similarity Reasoning

Regarding features of visual and textual scene graphs, CFSGR conducts graph similarity reasoning by measuring the similarity between visual objects and textual words, visual predicates and textual triplets, visual and textual features.

Before local graph similarity calculation, CFSGR updates the features of the visual scene graph by leveraging the GCN (Graph Convolutional Network) [9] . Concretely, given the features of visual scene graph $SG_I = (O_I, R_I)$ (visual objects $emb(o_I)$ and predicates $emb(r_I)$, CFSGR uses a single-layer GCN to implement message pass for visual object and predicate features, a single-layer updating of GCN can be represented as:

$$\begin{aligned} emb(o_{Ii})' &= g_o^I\big(emb(o_{Ii})\big) \\ emb(r_{Ik})' &= g_r^I\left(emb(o_{Ii}), emb(r_{Ik}), emb\big(o_{Ij}\big)\right) \end{aligned} \tag{3}$$

The g_o and g_r are fully-connected layers with tanh activation functions. Given the updated visual objects $emb(o_I)'$ and textual words $emb(w_T)$, we follow SGM [15] to calculate their$N_w \times N_o$score matrix by $emb(w_T)^T emb(o_I)'$, shown as the blue matrix in Figure 1. Then, CFSGR implements max pooling for each row of score matrix, and averages them as the object-word score of visual and textual scene graphs, which can be represented as:

$$S_o^{local}(I,T) = \frac{1}{N_w}\sum_{j=1}^{N_w} max_{i\in[1,N_o]} emb\left(w_{Tj}\right)^T emb(o_{Ii})' \tag{4}$$

The N_o and N_w separately denote the numbers of visual objects and textual words. Similarly, given the updated visual predicates $emb(r_{Ik})'$ and textual triplets $emb(t_{Tk})$, CFSGR calculates their $N_T \times N_o$ score matrix by $emb(t_T)^T emb(r_I)'$, which is represented as the green matrix in Figure 1. Then, CFSGR implements max-pooling and averaging operations for each row of score matrix:

$$S_p^{local}(I,T) = \frac{1}{N_t}\sum_{j=1}^{N_t} \max_{i\in[1,N_r]} emb\left(t_{Tj}\right)^T emb(r_{Ii})' \tag{5}$$

The N_r and N_t separately denote the numbers of visual predicates and textual triplets. Besides calculating the local graph similarity for word-object pairs and predicate-triplet pairs, CFSGR conducts comprehensive graph similarity reasoning for all the nodes in visual and textual scene graphs. Concretely, CFSGR firstly calculates the global features for concatenated visual local features $concat\left(emb(o_I), emb(r_I)\right)$ and concatenated textual local features $concat\left(emb(w_T), emb(t_T)\right)$ through their corresponding self-attention networks, the calculation of which is represented as Eq. . The visual local features and visual global feature are respectively represented as $f_v^l = concat\left(emb(o_I), emb(r_I)\right)$ and f_v^g, while the textual local features and textual global feature are respectively represented as $f_t^l = concat\left(emb(w_T), emb(t_T)\right)$ and f_t^g. For visual local features, CFSGR calculates the textual-attended representations of visual scene graph through SCAN [10] attention:

$$f_v^{ta} = scan\_attention(f_t^l, f_v^l) \tag{6}$$

The f_t^l serves as queries to calculate the attention weights for f_v^l by matrix multiplication. Then, the graph similarity representation r for comprehensive graph similarity reasoning can be obtained through:

$$\begin{aligned} r^l &= l_2norm((f_t^l - f_v^{ta})^2) \\ r^g &= l_2norm\left(\left(f_t^g - f_v^g\right)^2\right) \\ r &= concat(r^g, r^l) \end{aligned} \tag{7}$$

The $l_2norm(*)$ denotes the l_2 normalization function, $concat(*)$ represents the concatenation operation. Finally, the comprehensive graph similarity is secured by graph similarity reasoning module, whose calculation can be represented as:

$$\begin{aligned} r_q &= MLP_q(r) \\ r_k &= MLP_k(r) \\ r_k &= MLP_k(r) \\ w &= \delta\left(r_k^T r_q\right) \\ r_s &= w \times r \\ S^{graph} &= \sigma\left(\left(\phi\left(MLP_s(r_s)\right)\right)\right) \end{aligned} \tag{8}$$

The $\delta(*)$denotes the Softmax activation function, $\phi(*)$ denotes the ReLU activation function, $\sigma(*)$denotes the Sigmoid function, $MLP_*(*)$ refers to the Multilayer Perceptron. The coarse-grained similarity between visual and textual scene graphs is the combination of local graph similarity and comprehensive graph similarity:

$$Sim_C = S_o^{local} + S_p^{local} + S^{graph} \tag{9}$$

3.2 Fine-grained Object and Predicate Similarity Reasoning

Same to the updating of visual scene graph in Eq. (3), for textual scene graph $SG_T = (O_T, R_T)$, CFSGR also obtains its updated features of objects and predicates from triplet-order paths $emb(t_T)$ through a single layer GCN:

$$\begin{aligned} emb(t_T) &= \left(emb(o_{Ts}), emb(r_T), emb(o_{Tk})\right) \\ emb(o_{Ti})' &= g_o^T\left(emb(o_{Ti})\right) \\ emb(r_{Tk})' &= g_r^T\left(emb(o_{Ti}), emb(r_{Tk}), emb\left(o_{Tj}\right)\right) \end{aligned} \tag{10}$$

As shown in Figure 2, regarding updated features of objects and predicates in visual and textual scene graphs, CFSGR firstly calculates their corresponding global representations through self-attention network (for objects) and conv-attention network (for predicates). The calculations of the self-attention network can be represented as:

$$\begin{aligned} emb\left(o_q\right) &= W_o\phi\left(AvePool(emb(o)')\right) \\ a_o &= \sigma((emb(o)'^T emb\left(o_q\right)) \\ emb(o^g) &= AvePool(a_o^T \odot emb(o)') \end{aligned} \tag{11}$$

The $\phi(*)$ denotes the ReLU activation function, $\sigma(*)$ denotes the Sigmoid function, $AvePool(*)$ refers to the average pooling operation. The calculations of conv-attention network are represented as:

$$\begin{aligned} emb(r_d) &= conv1d_d(emb(r)') \\ emb(r^g) &= MaxPool\left(conv1d_u\left(emb(r_d)\right)\right) \end{aligned} \tag{12}$$

The $conv1d_d(*)$ and $conv1d_u(*)$ separately denote the 1d convolutional layers for scaling down and up the inputs, $MaxPool(*)$ refers to the max pooling operation.

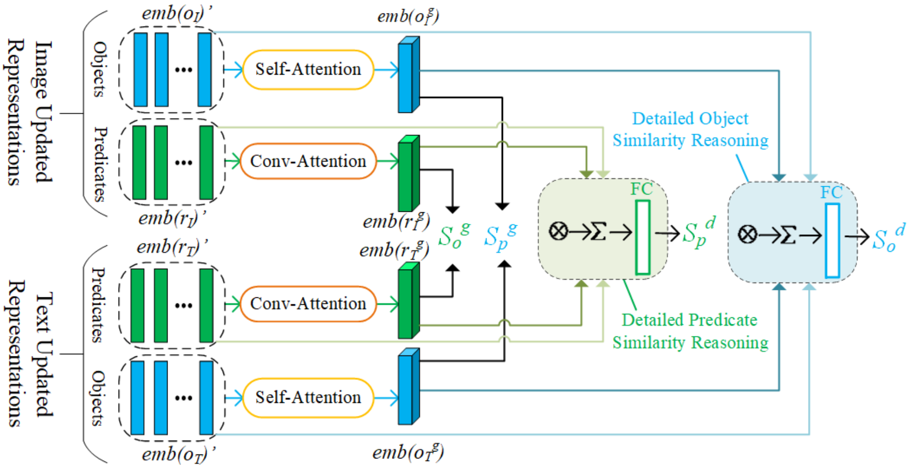

Fig. 2. Fine-grained object and predicate similarity reasoning. Regarding visual updated features $emb(o_I)', emb(r_I)'$ and textual updated features $emb(o_T)', emb(r_T)'$, CFSGR firstly obtains global object representations $emb\left(o_I^g\right), emb\left(o_T^g\right)$ and global predicate representations $emb\left(r_I^g\right), emb\left(r_I^g\right)$. Then, CFSGR calculates the cosine similarity of objects and predicates with their corresponding global representations. Besides, CFSGR conducts detailed similarity reasoning for objects and predicates based on their local features$emb(o)', emb(r)'$ and global representations $emb(o^g), emb(r^g)$, separately.

Regarding global object representations $emb\left(o_v^g\right)$ and $emb\left(o_t^g\right)$, CFSGR takes their cosine similarity as global object similarity:

$$S_o^g = \frac{abs\left(emb\left(o_I^g\right) - emb\left(o_T^g\right)\right)}{||emb\left(o_I^g\right) - emb\left(o_T^g\right)||_1} \tag{13}$$

Where $abs(*)$ obtains the absolute value of inputs. Similarly, for global predicate representations $emb\left(r_v^g\right)$ and $emb\left(r_t^g\right)$, CFSGR takes their cosine similarity as global object similarity:

$$S_r^g = \frac{abs\left(emb\left(r_I^g\right) - emb\left(r_T^g\right)\right)}{||emb\left(r_I^g\right) - emb\left(r_T^g\right)||_1} \tag{14}$$

Besides cosine similarity calculation for global representations, CFSGR also separately performs detailed similarity reasoning for object similarity representation r_o and predicate similarity representation r_p. Same to graph representation r calculation in Eq. (7),r_o and r_p are calculated as Eq. (15):

$$
\begin{aligned}
r_o^l &= l_2norm(emb(o_I)' - emb(o_T)') \\
r_o^g &= l_2norm\left(emb\left(o_I^g\right) - emb\left(o_T^g\right)\right) \\
r_o &= concat\left(r_o^g, r_o^l\right) \\
r_p^l &= l_2norm(emb(r_I)' - emb(r_T)') \\
r_p^g &= l_2norm\left(emb\left(r_I^g\right) - emb\left(r_T^g\right)\right) \\
r_p &= concat\left(r_p^g, r_p^l\right)
\end{aligned}
\tag{15}
$$

Finally, the detailed object similarity and predicate similarity are calculated as Eq. (16):

$$
\begin{aligned}
r_o^{attn} &= l_1norm\left(\sigma\left(BN\left(MLP(r_o)\right)\right)\right) \\
S_o^d &= l_2norm((r_o^{attn})^T r_o) \\
r_p^{attn} &= l_1norm\left(\sigma\left(BN\left(MLP\left(r_p\right)\right)\right)\right) \\
S_p^d &= l_2norm\left(\left(r_p^{attn}\right)^T r_p\right)
\end{aligned}
\tag{16}
$$

The $l_1norm(*)$ denotes thel_1 normalization function, while the $BN(*)$ refers to the batch normalization function. The fine-grained similarity of objects and predicates in visual and textual scene graphs is the combination of their cosine similarity and detailed similarity:

$$
Sim_F = S_o^g + S_p^g + S_o^d + S_p^d \tag{17}
$$

3.3 Loss function for training the CFSGR

CFSGR simply uses hinge-based triplet ranking loss with hard negative mining loss function [3] $Loss_H$, which is widely adopted by existing image-text retrieval methods, to individually train the coarse-grained and fine-grained similarity reasoning of CFSGR, as shown in Eq. (18):

$$
\begin{aligned}
Loss_H = max&\left(0, m - Sim_*(I,T) + Sim_*\left(I,\overline{T}\right)\right) \\
+max&\left(0, m - Sim_*(I,T) + Sim_*\left(\overline{I},T\right)\right)
\end{aligned}
\tag{18}
$$

The m denotes the margin parameter, $Sim(I,T)$ denotes the similarity score for matched image-text pairs in the mini-batch, $\overline{I}$ and $\overline{T}$ refers to the hardest negative image and text of T and I in the mini-batch, respectively.

4 Experiments

4.1 Dataset and Metrics

To achieve a comprehensive evaluation, CFSGR is evaluated on a commonly used retrieval dataset, Flickr30K [19] , which contains $31{,}000$ images. Each image in both datasets has 5 matched captions. CFSGR takes $29{,}000$ images for training, $1{,}000$ images for validation, $1{,}000$ images for testing.

The proposed CFSGR and methods for comparison are evaluated with $Recall@K(R@K)$, which denotes the proportion of query $image/text$ whose cross-modal counterparts are in the top-K ranking results. The K is set as 1, 5, and 10. In addition, the overall evaluation metric $Rsum$ is defined as Eq. (19):

$$Rsum = \underbrace{R@1 + R@5 + R10}_{Text\ etrieval} + \underbrace{R@1 + R@5 + R10}_{Image\ etrieval} \tag{19}$$

4.2 Environment and Setup

The CFSGR framework is implemented on the PyTorch platform, with all experiments conducted on a single Nvidia RTX 4090 GPU. Following LGSGM [11], visual object and predicate features are extracted using EfficientNet-b5[14]. The initial learning rates for both the coarse-grained graph similarity reasoning network and the fine-grained similarity reasoning network are set to $4e-4$. The training epochs are configured as 45 for the coarse-grained module and 60 for the fine-grained module.

4.3 Results and Analysis

CFSGR picks several SOTA works at their release moments from object feature based methods and scene graph based methods for comparison, as introduced in Section 2. For object feature based methods for comparison, they are SHAN [8], VSE∞ [3], DCPA [13], WCGL [17], SGRAF [5] and CGMN [4]. For scene graph based methods for comparison, they are SGM [5], LGSGM [11], HSGMP [6], SSAMT [7] and SGSIN [12].

Table 1 reports the bidirectional retrieval results of CFSGR and methods for comparison on Flickr30K [19]. It can be observed that the CFSGR outperforms all the methods with the best $Rsum$ of 506. Among scene graph based methods, CFSGR achieves the best performance in 6 out of 7 metrics. In image retrieval, CFSGR demonstrates notable improvements across multiple metrics compared to both object feature based and scene graph based methods.

When compared with CGMN [4] and LGSGM [11], two methods that achieve the best performance in their respective categories, CFSGR attains a value of 61.8 in $R@1$ of image retrieval. This is 1.9 points higher than CGMN [4] (59.9) and 4.4 points higher than LGSGM [11] (57.4), as highlighted in Table 1. For $R@5$, CFSGR scores 86.4, outperforming CGMN [4] (85.1) by 1.3 points and LGSGM [11] (84.1) by 2.3 points. In $R@10$, CFSGR reaches 92.1, surpassing CGMN [4] (90.6) by 1.5 points and

LGSGM [11] (90.2) by 1.9 points. These results highlight CFSGR's consistent and significant performance improvements in image retrieval on Flickr30K [19].

Table 1. Methods' performance on Flickr30K [19]. The ***bold*** results of each metric denotes the best among scene graph based methods. The *underlined* results denote the best among all methods.

Model	Text retrieval			Image retrieval			$Rsum$
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
Object Feature based Method							
$SHAN$	74.6	93.5	96.9	55.3	81.3	88.4	490.0
$VSE\infty$	76.5	94.2	97.7	56.4	83.4	89.9	498.1
$DCPA$	75.3	93.7	97.5	59.9	85.0	91.9	502.5
$WCGL$	74.8	93.3	96.8	54.8	80.6	87.5	487.8
$SGRAF$	77.8	94.1	97.4	58.5	83.0	88.8	499.6
$CGMN$	77.9	93.8	96.8	59.9	85.1	90.6	504.1
Scene Graph based Methods							
SGM	71.8	91.7	95.5	53.5	79.6	86.5	478.6
$LGSGM$	71.0	91.9	96.1	57.4	84.1	90.2	490.7
$HSGMP$	73.4	93.0	96.8	55.0	81.4	88.2	487.8
$SSAMT$	75.4	92.6	96.4	54.8	81.5	88.0	488.7
$SGSIN$	73.1	**93.6**	96.8	53.9	80.1	87.2	484.0
$CFSGR(ours)$	**75.1**	93.4	**97.2**	**61.8**	**86.4**	**92.1**	**506.0**

However, in text retrieval, CFSGR achieves the suboptimal among all methods and is slightly improved when compared with the scene graph based methods, The authors assume that the reason is the textual scene graphs built upon similar texts have close local and global vectors in the feature space, the similarities of matched pairs and unmatched pairs are close to each other. Therefore, this phenomenon eventually results in weak improvements of text retrieval.

4.4 Ablation Study

The ablation study in Table 2 evaluates the contribution of main components in the CFSGR for image-text retrieval on Flickr30K [19]. The study examines the impact of three key components: coarse-grained graph similarity reasoning (coarse), fine-grained object similarity reasoning (fine-obj) and fine-grained predicate similarity reasoning (fine-pre). The full model, CFSGR, integrates all three components, and its performance is compared against variants that exclude one of these components.

In terms of $Rsum$, CFSGR scores the highest 506.0, outperforming all ablated and single-module networks. This highlights the synergistic effect of combining coarse-grained and fine-grained reasoning.

By comparing the performance of CFSGR with all ablated networks, it is observed that removing fine-pre has the least impact on the performance degradation of CFSGR. The authors attribute this to the limited number of textual triples extracted from the dataset, which results in fewer predicate elements in the textual scene graphs, thereby diminishing the role of fine-pre. This reason is further supported by the fact that fine-pre scores the lowest $Rsum$ (487.1) compared to the other two single-module networks, coarse (489.3) and fine-obj (487.1).

Table 2. Ablation study for CFSGR on Flickr30K [19]. The best results of each metric are in ***bold***. 'coarse' denotes the coarse-grained graph similarity reasoning network, 'fine-obj' denotes the fine-grained object similarity reasoning network, 'fine-pre' denotes the fine-grained predicate similarity reasoning network. 'w/o' refers to without.

Model	Text retrieval			Image retrieval			$Rsum$
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
$coarse$	69.3	92.0	95.7	58.2	84.1	90.0	489.3
$fine-obj$	69.9	92.3	96.2	57.1	83.3	89.7	488.5
$fine-pre$	70.7	90.5	95.4	57.0	83.5	90.0	487.1
$CFSGR\ w/o\ coarse$	73.7	93.9	96.3	60.2	85.6	91.2	500.0
$CFSGR\ w/o\ fine-obj$	72.5	93.4	96.8	61.0	85.6	91.4	500.7
$CFSGR\ w/o\ fine-pre$	72.8	**93.7**	97.1	60.8	85.6	91.4	501.4
$CFSGR$	**75.1**	93.4	**97.2**	**61.8**	**86.4**	**92.1**	**506.0**

5 Conclusion

In this paper, we introduce the Coarse-to-Fine Scene Graph Similarity Reasoning (CFSGR) method, a novel framework designed to enhance image-text retrieval by separately conducting coarse-grained and fine-grained cross-modal similarity reasoning. CFSGR devises local and global alignments at different granularities, ensuring a comprehensive utilization of scene graph features. The coarse-grained network focuses on graph-level similarity reasoning, while the fine-grained network aligns objects and predicates, enabling a more detailed and effective cross-modal matching.

Experimental results on the benchmark dataset Flickr30K demonstrate the superiority of CFSGR over SOTA methods, achieving a competitive $Rsum$ score of 506. The ablation study further validates the effectiveness of each component, highlighting the synergistic impact of combining coarse-grained and fine-grained reasoning. Future work will explore enhancing the predicate reasoning component and extending the framework to other multi-modal retrieval tasks.

Acknowledgments. This work is supported by the National Key R&D Program of China (2023YFC3304903).

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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, pp. 6077–6086 (2018)

2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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. pp. 6077–6086 (2018)
3. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15789–15798 (2021)
4. Cheng, Y., Zhu, X., Qian, J., Wen, F., Liu, P.: Cross-modal graph matching net-work for image-text retrieval. ACM Transactions on Multimedia Computing, Com-munications, and Applications (TOMM) 18(4), 1–23 (2022)
5. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1218–1226 (2021)
6. Duan, Y., Xiong, Y., Zhang, Y., Fu, Y., Zhu, Y.: Hsgmp: Heterogeneous scene graph message passing for cross-modal retrieval. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. pp. 82–91 (2021)
7. Fan, Z., Wei, Z., Li, Z., Wang, S., Shan, H., Huang, X., Fan, J.: Constructing phrase-level semantic labels to form multi-grained supervision for image-text re- trieval. In: Proceedings of the 2022 International Conference on Multimedia Re-trieval. pp. 137–145 (2022)
8. Ji, Z., Chen, K., Wang, H.: Step-wise hierarchical alignment network for image-text matching. arXiv preprint arXiv:2106.06509 (2021)
9. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
10. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV). pp. 201–216 (2018)
11. Nguyen, M.D., Nguyen, B.T., Gurrin, C.: A deep local and global scene-graph matching for image-text retrieval. In: New Trends in Intelligent Software Method-ologies, Tools and Techniques, pp. 510–523 (2021)
12. Pei, J., Zhong, K., Yu, Z., Wang, L., Lakshmanna, K.: Scene graph semantic infer-ence for image and text matching. ACM Transactions on Asian and Low-Resource Language Information Processing 22(5), 1–23 (2023)
13. Shi, Z., Zhang, T., Wei, X., Wu, F., Zhang, Y.: Decoupled cross-modal phrase-attention network for image-sentence matching. IEEE Transactions on Image Pro-cessing 33, 1326–1337 (2022)
14. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
15. Wang, S., Wang, R., Yao, Z., Shan, S., Chen, X.: Cross-modal scene graph match-ing for relationship-aware image-text retrieval. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1508–1517 (2020)
16. Wang, Y., Yasunaga, M., Ren, H., Wada, S., Leskovec, J.: Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21582–21592 (2023)
17. Wang, Y., Zhang, T., Zhang, X., Cui, Z., Huang, Y., Shen, P., Li, S., Yang, J.: Wasserstein coupled graph learning for cross-modal retrieval. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1793–1802. IEEE (2021)
18. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10685–10694 (2019)

19. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the association for computational linguistics 2, 67–78 (2014)
20. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5831–5840 (2018)