

MKVSE: Multimodal Knowledge Enhanced Visual-Semantic Embedding for Image-Text Retrieval

DUODUO FENG, Peking University, China

XIANGTENG HE, Peking University, China

YUXIN PENG*, Peking University, China and Peng Cheng Laboratory, China

Image-text retrieval aims to take the text (image) query to retrieve the semantically relevant images (texts), which is fundamental and critical in the search system, online shopping, and social network. Existing works have shown the effectiveness of visual-semantic embedding and unimodal knowledge exploiting (e.g. textual knowledge) in connecting the image and text. However, they neglect the implicit multimodal knowledge relations between these two modalities when the image contains information that is not directly described in the text, hindering the ability to connect the image and text with the implicit semantic relations. For instance, an image shows a person next to the "tap" but the pairing text description may only include the word "wash", missing the washing tool "tap". The implicit semantic relation between image object "tap" and text word "wash" can help to connect the above image and text. To sufficiently utilize the implicit multimodal knowledge relations, we propose a **Multimodal Knowledge Enhanced Visual-Semantic Embedding (MKVSE)** approach building a multimodal knowledge graph to explicitly represent the implicit multimodal knowledge relations and injecting it to visual-semantic embedding for image-text retrieval task. The contributions in this paper can be summarized as follows: (1) **Multimodal Knowledge Graph (MKG)** is proposed to explicitly represent the implicit multimodal knowledge relations between the image and text as *intra-modal semantic relations* and *inter-modal co-occurrence relations*. Intra-modal semantic relations provide synonymy information that is implicit in the unimodal data such as the text corpus. And inter-modal co-occurrence relations characterize the co-occurrence correlations (such as temporal, causal, and logical) which are implicit in image-text pairs. These two relations help establishing reliable image-text connections in the higher-level semantic space. (2) **Multimodal Graph Convolution Networks (MGCN)** is proposed to reason on the MKG in two steps to sufficiently utilize the implicit multimodal knowledge relations. In the first step, MGCN focuses on the intra-modal relations to distinguish other entities in the semantic space. In the second step, MGCN focuses on the inter-modal relations to connect multimodal entities based on co-occurrence correlations. The two-step reasoning manner can sufficiently utilize the implicit semantic relations between two modal entities to enhance the embeddings of the image and text. Extensive experiments are conducted on two widely-used datasets, namely Flickr30K and MSCOCO, to demonstrate the superiority of the proposed MKVSE approach in achieving state-of-the-art performances. The codes are available at <https://github.com/PKU-ICST-MIPL/MKVSE-TOMM2023>.

CCS Concepts: • **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Knowledge representation and reasoning**.

*Corresponding author.

This work was supported by the grants from the National Natural Science Foundation of China (62132001, 61925201, 62272013, U22B2048) and by 2022 Tencent Wechat Rhino-Bird Focused Research Program.

Authors' addresses: Duoduo Feng, Peking University, Wangxuan Institute of Computer Technology, Beijing, 100871, China; Xiangteng He, Peking University, Wangxuan Institute of Computer Technology, Beijing, 100871, China; Yuxin Peng, Peking University, Wangxuan Institute of Computer Technology, Beijing, 100871, China and Peng Cheng Laboratory, Shenzhen, 518055, China, pengyuxin@pku.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

1551-6857/2023/January-ART1 \$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

Additional Key Words and Phrases: Image-Text Retrieval, Cross-Modal Retrieval, Visual-Semantic Embedding, Multimodal Knowledge Graph

ACM Reference Format:

Duoduo Feng, Xiangteng He, and Yuxin Peng. 2023. MKVSE: Multimodal Knowledge Enhanced Visual-Semantic Embedding for Image-Text Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, Article 1 (January 2023), 21 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

The image and text are the two most prevalent modalities for understanding the real world. Correspondingly, the demands for effective and efficient image-text retrieval technologies are significantly increasing, which is a fundamental and critical problem in multimodal retrieval and has attracted extensive attention in recent years [7, 14, 36, 37, 46, 52]. Specifically, it aims to retrieve the texts (images) that are most semantically relevant to the given image (text) query. However, textual descriptions are abstract while visual scenes are specific which contain redundant information. They exhibit heterogeneous properties with inconsistent distributions and representations, making it quite challenging to measure the semantic similarity between the image and text.

To tackle this challenge, Visual-Semantic Embedding (VSE) [7, 13, 14] was proposed to learn a unified joint embedding space, where the similarities between the embeddings of paired image and text entities are optimized to be maximum. A large proportion of methods utilize the deep neural networks to extract the global representations of both images and texts, and then learn to measure the similarity by some criterion. Wang et al. [47] proposed a two-branch neural network to extract the global embeddings of images and texts respectively, and then fuse the two branches via element-wise product for learning the similarity between these two data modalities. Faghri et al. [14] proposed a loss function using hard negative mining to improve the quality of VSE models by learning with a maximized hinge-based triplet ranking loss. Wang et al. [48] proposed to represent image and text with scene graphs: visual scene graph and textual scene graph, each of which is exploited to jointly characterize objects and relationships in the corresponding modality. Then the image-text retrieval task is naturally formulated as cross-modal scene graph matching. However, this paradigm *neglects any prior unimodal knowledge*, which may hinder its capabilities to reason the knowledge relations between image and text. To address this problem, some works incorporate the commonsense knowledge by exploiting the word co-occurrences for reasoning the high-level relations between image and text. Wang et al. [46] proposed a consensus-aware visual-semantic embedding method to exploit the consensus information from the image captioning corpus. It computes co-occurrences between the word concepts and learning the consensus-aware-concept representations for image-text retrieval. Shi et al. [41] proposed a scene concept graph to incorporate scene knowledge by utilizing the objects frequently appearing in the same scene. Semantic concepts are detected from images and then expanded by the scene concept graph to select relevant contextual concepts and fuse their representations with the image embedding feature. However, these methods only *utilize unimodal knowledge* (e.g. textual knowledge) and *neglect the implicit multimodal knowledge relations between the image and text*. When the image contains information that is not directly described in the text, the implicit multimodal knowledge relations can help to connect the image and text in the higher-level semantic space.

To tackle the aforementioned issues, we propose the **Multimodal Knowledge enhanced Visual-Semantic Embedding (MKVSE)** approach. As illustrated in Figure 1, the multimodal knowledge graph is built to explicitly represent the implicit multimodal relations. And then it can be used to support the downstream image-text retrieval for connecting the image and text. When we see the words "washing vegetables", we associate a image of the water, the tap, etc. When we see the words

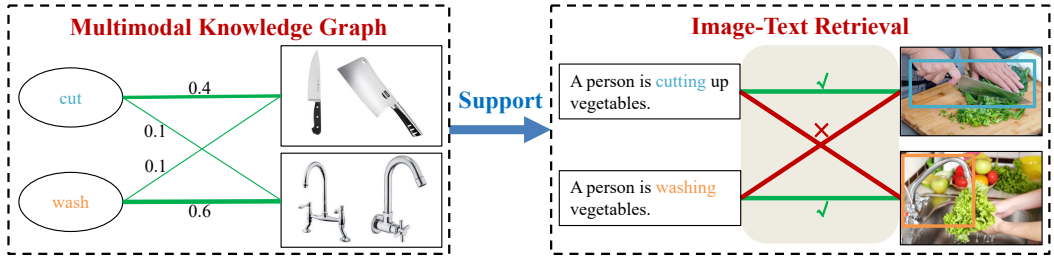


Fig. 1. The illustration of the multimodal knowledge graph supporting image-text retrieval.

"cutting up vegetables", we picture a image with the knife and vegetables in our mind. The main contributions of this paper can be summarized as follows:

- **Multimodal Knowledge Graph (MKG)** is proposed to explicitly represent implicit multimodal knowledge relations between the image and text, including both image and text entities connected by their intra-modal semantic relations and inter-modal co-occurrence relations. Intra-modal semantic relations provide synonymy information that is implicit in the unimodal data such as the text corpus. And inter-modal co-occurrence relations characterize the co-occurrence correlations which are implicit in image-text pairs. These two relations can help to connect the image and text in the higher-level semantic space
- **Multimodal Graph Convolution Networks (MGCN)** is proposed to sufficiently utilize the implicit multimodal knowledge by reasoning on the MKG in two steps. MGCN can focus on different aspects in each step. Specifically, in the first step, MGCN separately reasons on the image entities and text entities to focus on intra-modal semantic relations. It aims to distinguish other entities in the semantic space. In the second step, MGCN reasons on the whole multimodal graph to focus on the inter-modal co-occurrence relations. It aims to connect multimodal entities based on statistic correlations. Finally, the implicit semantic relations between two modal entities can be mined to enhance the embeddings of the image and text for a better retrieval performance.

Extensive experiments on two widely-used datasets, namely Flickr30K [56] and MSCOCO [28], are conducted. The experimental results demonstrate the effectiveness and superiority of our proposed MKVSE approach.

The rest of this paper is organized as follows: Section 2 summarizes a brief review of related works. Section 3 introduces the proposed MKVSE approach and explains its architecture in detail. Section 4 presents the experiments and analyses, including comparison with state-of-the-art image-text retrieval methods, ablation study, and visualization results. Section 5 concludes this paper and presents the future work.

2 RELATED WORK

2.1 Image-Text Retrieval

Measuring the image-text semantic similarity is essential for image-text retrieval. Based on how the similarity is measured, existing image-text retrieval methods can be roughly categorized into two groups: cross-interaction matching methods [8, 20, 23, 27, 33, 50] and independent representation matching methods [7, 14, 15, 24, 43, 46]. As for cross-interaction matching methods, Karpathy et al. [20] first adopted R-CNN [17] to detect salient objects and inferred latent alignments between word-level textual features in sentences and region-level visual features in images. Moreover, an

cross-attention mechanism [8, 23, 27, 33, 50] was applied to capture the fine-grained interaction between images and texts for the image-text retrieval. Although achieving high performance, cross-attention suffers from calculation explosion during inference for requiring to forward over all pairs of images and texts, which cannot be ignored in retrieval tasks [7]. Compared with cross-interaction matching methods, independent representation matching methods are much more efficient during inference. Frome et al. [15] proposed a pioneer embedding based method which projected image features and skip-gram word features by a linear mapping and calculated similarity accordingly. Faghri et al. [14] proposed VSE++ to further improve the quality of Visual-Semantic Embeddings (VSE) by learning with online hard-negative mining. Beyond the above, more researches along this line focus on improving the visual or text representation, or designing auxiliary training objectives [7, 24, 43]. Li et al. [24] proposed Visual Semantic Reasoning Network (VSRN) to address the lack of global semantic concepts in the current representation of the image, which is in the image's corresponding text caption. VSRN generates enhanced visual representations by capturing key objects and semantic concepts of a scene. To handle polysemous entities with multiple possible meanings, Song et al. [43] proposed Polysemous Visual-Semantic Embedding (PVSE) to compute multiple and diverse representations of an entity by combining global context with locally-guided features. The two polysemous instance embedding networks are tied up and optimized jointly in the multiple instance learning framework. Chen et al. [7] proposed Generalized Pooling Operator (GPO) to automatically seek the best pooling function for different data modality and feature extractor, requiring no manual tuning while staying effective and efficient. However, all the above methods only rely on the image-text pairs, neglecting the prior knowledge between the the image and text. Shi et al. [41] built a Scene Concept Graph (SCG) by considering co-occurrence pairs of semantic concepts in the scene graph of images. Co-occurred concepts in the same scene can provide common-sense knowledge to discover other semantic-related concepts. Then the SCG can be used to expand more semantic concepts for enhancing image representation semantically. Wang et al. [46] proposed a Consensus-aware Visual-Semantic Embedding (CVSE) model to incorporate the consensus information into image-text matching. And the consensus information is exploited by computing the statistical co-occurrence correlations between the semantic concepts from the image captioning corpus. However, these methods only utilize the unimodal knowledge (such as knowledge from the scene graphs of images and the image captions of text corpus) and neglect the implicit multimodal knowledge relations between the image and text when the image contains information that is not directly described in the text which hinders the ability of connecting image and text. Hence exploring the effectiveness of implicit multimodal knowledge relations for image-text retrieval is necessary. Our proposed approach introduces the multimodal knowledge graph to explicitly represent implicit multimodal knowledge relations between the image and text, which can help to connect two modalities in the higher-level semantic space.

2.2 Multimodal Graph Based Deep Learning

Multimodal graph based deep learning can take advantage of the multimodal content for various multimodal understanding tasks, such as fake news detection [42], video emotion recognition [12], multimodal neural machine translation [19], recommendation systems [59] and image-text retrieval [16]. Wang et al. [49] proposed an end-to-end knowledge-driven multimodal graph convolution network to model the semantic-level representations for fake news detection by jointly modeling the textual information, knowledge concepts and visual information into a unified deep model. Mai et al. [31] proposed a hierarchical graph fusion network which can explicitly model unimodal, bimodal and trimodal dynamics for video emotion recognition. Yin et al. [55] proposed a graph-based multi-modal fusion encoder to exploit fine-grained semantic correspondences between semantic units of different modalities for multi-modal neural machine translation. Sun et al. [44] incorporated

multi-modal knowledge graph and employs information propagation on it to obtain better entity embeddings for recommendation. Moreover, some works [9, 16, 48] also introduced the multimodal graph to the image-text retrieval. Garcia et al. [16] proposed to enhance visual representations from neural networks with contextual artistic information. To this end, an art-specific knowledge graph is built to capture contextual relations between artistic attributes, which is used to inform the visual model. However, the multimodal knowledge graph for all candidate images needs building before retrieval, which may hinder its application when the candidate images lack contextual relations (such as their authors and dates). Wang et al. [48] proposed to represent image and text with two kinds of scene graphs: visual and textual scene graph, each of which is exploited to jointly characterize objects and relations in the two modalities. The image-text retrieval task is then naturally formulated as cross-modal scene graph matching. Cheng et al. [9] proposed a graph-based Cross-modal Graph Matching Network (CGMN) to explore both intra-relations and inter-relations without introducing network interaction. CGMN can take the advantages of cross-modal inter-relation reasoning while being as efficient as the independent methods. However, the multimodal graphs in the above methods are all built while training upon the image-text pairs, on which the implicit multimodal knowledge relations are not represented explicitly. Thus, they cannot sufficiently utilize the implicit multimodal relations in the multimodal knowledge graph, which is crucial for image-text retrieval. Our proposed approach introduces the multimodal knowledge graph to explicitly represent implicit multimodal knowledge relations between the image and text, and then the proposed multimodal graph convolution networks can sufficiently utilize the implicit multimodal knowledge relations in a two-step manner. In each step, it can focus on different aspects to further improve the quality of visual-semantic embedding.

2.3 Multimodal Knowledge Enhanced Deep Learning

Although deep learning has achieved great success in many fields, it is usually data hungry, lacks of interpretability, and fails to perform well on unseen situations. Various kinds of prior knowledge often exist in the target domain, which can alleviate the deficiencies of deep learning. The existing multimodal knowledge enhanced deep learning methods [11, 53, 58] aim to incorporate multimodal knowledge into the networks, which has been utilized in various multimodal understanding tasks, such as visual question answering [11], video caption [58], multimodal named entity recognition [5], dialogue systems [54], and image-text retrieval [34, 53]. Ding et al. [11] proposed to represent multimodal knowledge by the form of triplets to correlate visual objects and fact answers, which constructs vision-relevant multimodal knowledge for the VQA scenario. Zhang et al. [58] proposed knowledge-enhanced spatial-temporal inference on product-oriented spatial-temporal graphs to capture the dynamic change of fine-grained product-part characteristics. Chen et al. [5] proposed to introduce external multi-modal knowledge helping improve named entity extraction, where the matched entity information is incorporated into the model for feature fusion. Yang et al. [54] proposed to integrate external multimodal knowledge base reasoning with pre-trained language models on task-oriented dialogue systems, which enhances the model via a multi-granularity fusion mechanism to capture multi-grained semantics in the dialogue history. Yang et al. [53] proposed to perform triple contrastive learning in pre-training, which takes the advantage of localized and structural information from image and text input to benefits in representation learning. Nian et al. [34] proposed a multi-modal knowledge representation learning framework which attempts to handle knowledge from both textual and visual modal web data. However, these methods neglect the multimodal implicit relations when the image contains objects that are not directly described in the text, which hinders the ability to connect the image and the text. Our proposed approach can explicitly represent the implicit multimodal knowledge relations between the image and the text, and then it can be integrated to the network for more robust image-text connections.

1:6

Feng et al.

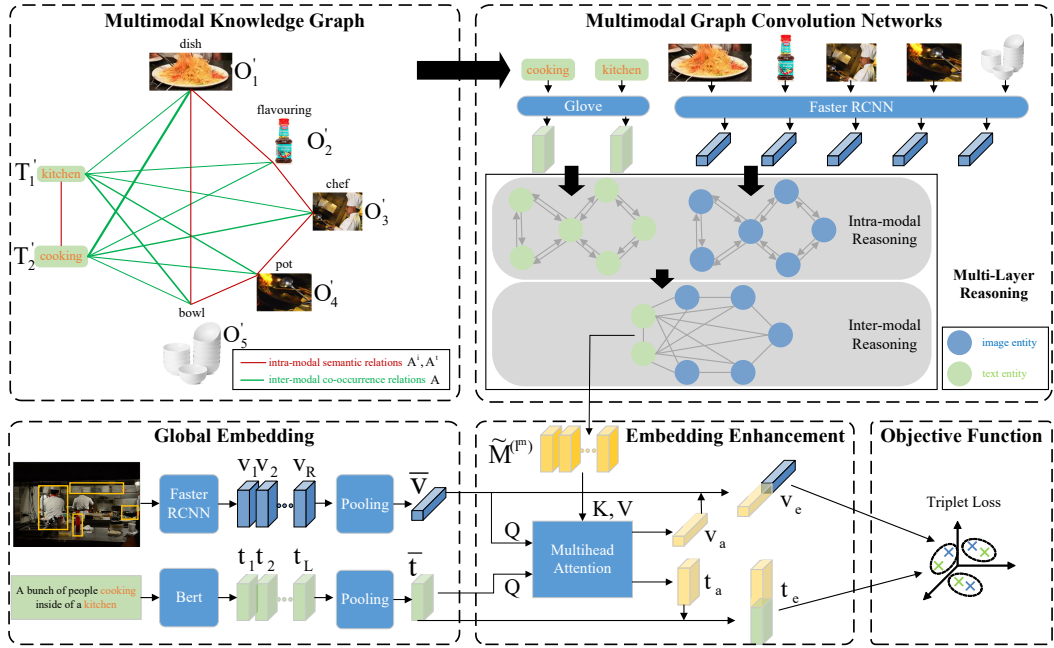


Fig. 2. Overview of the proposed MKVSE framework.

3 MULTIMODAL KNOWLEDGE ENHANCED VISUAL-SEMANTIC EMBEDDING

As shown in Figure 2, our proposed Multimodal Knowledge Enhanced Visual-Semantic Embedding approach (MKVSE) comprises five components: Global Embedding (Section 3.1), Multimodal Knowledge Graph (Section 3.2), Multimodal Graph Convolution Networks (Section 3.3), Embedding Enhancement (Section 3.4), and Objective Function (Section 3.5). In Global Embedding, each input image is represented as region-level features $V = [v_1; \dots; v_R]$, where v_i is the feature vector for the i th region [1]. Each input text is represented as word features $T = [t_1; \dots; t_L]$, where t_j is the feature vector of the j th word [45]. The pooling function GPO [7] is adopted to calculate the image global embedding \bar{v} from the image feature V and the text global embedding \bar{t} from the text feature T . In Multimodal Knowledge Graph (MKG), the entities $\{T'_1, T'_2, \dots, T'_{n^t}, O'_1, O'_2, \dots, O'_{n^i}\}$ of MKG are selected as the image objects appearing n^i th most frequently and the text words appearing n^t th most frequently in the Visual Genome dataset [22]. The relations of MKG are calculated as the co-occurrence relations A between two modalities and the semantic relations A^i, A^t within the modalities. The text entities are represented by GloVe embedding [35], and the image entities are represented as the mean pooling of the same category features [1]. In Multimodal Graph Convolution Networks (MGCN), MKG is reasoned by MGCN in a two-step manner, which can focus on different aspects in each step, to get the embeddings $\tilde{M}^{(lm)}$ of entities $\{T'_1, T'_2, \dots, T'_{n^t}, O'_1, O'_2, \dots, O'_{n^i}\}$ in MKG, where l_m is the number of MGCN's layers. In Embedding Enhancement, the entities' embeddings $\tilde{M}^{(lm)}$ are used to enhance the input image's global embedding \bar{v} and the input text's global embedding \bar{t} with multihead attention mechanism [45] to generate the enhanced embeddings v_e and t_e . In Objective Function, the enhanced embeddings v_e and t_e are aligned by optimizing the hinge-based bidirectional triplet loss. Finally, The candidate images (texts) of the top similarities with the text (image) query are selected as the final retrieval result.

3.1 Global Embedding

3.1.1 Global Embedding of Image. For an input image I , we follow [7, 36, 37] to detect salient regions with the Bottom-Up and Top-Down attention model (BUTD) [1], which selects the top R ($R = 36$) Region Of Interests (ROIs) with the highest class confidence scores. Then R region-level image features $\mathbf{V} = [\mathbf{v}_1; \dots; \mathbf{v}_R] \in \mathbb{R}^{R \times D_i}$, where D_i ($D_i = 2048$) is the dimension of the extracted region features. Afterwards, \mathbf{V} is projected into a D -dimensional space via a Fully-Connected (FC) linear projection. The obtained visual region representation is denoted as $\tilde{\mathbf{V}} = [\tilde{\mathbf{v}}_1; \dots; \tilde{\mathbf{v}}_R] \in \mathbb{R}^{R \times D}$. Moreover, we acquire the global embedding $\tilde{\mathbf{v}} \in \mathbb{R}^D$ of the input image I by adopting a pooling function on $\tilde{\mathbf{V}}$. Generalized Pooling Operator (GPO) [7] has achieved the impressive improvement in image-text retrieval compared with traditional pooling strategy such as mean-pooling and max-pooling, which is adopted as our pooling strategy.

3.1.2 Global Embedding of Text. For each input text T , we follow [7, 36, 37] to utilize pre-trained Bert [45] as the text encoder to extract word representation $\mathbf{T} = [\mathbf{t}_1; \dots; \mathbf{t}_L] \in \mathbb{R}^{L \times D_t}$, where $\mathbf{t}_j \in \mathbb{R}^{D_t}$ denotes the representation of T 's j th word, L denotes the number of words, and D_t denotes the dimension of word embedding. Then \mathbf{T} is projected into a D -dimensional space via a FC linear projection. The obtained textual word representation is denoted as $\tilde{\mathbf{T}} = [\tilde{\mathbf{t}}_1; \dots; \tilde{\mathbf{t}}_L] \in \mathbb{R}^{L \times D}$. The global embedding $\tilde{\mathbf{t}} \in \mathbb{R}^D$ of the input text T is acquired by adopting the same pooling function GPO [7] in Section 3.1.1.

3.2 Multimodal Knowledge Graph

To explicitly represent the implicit relations between the image and text, we build the Multimodal Knowledge Graph (MKG).

3.2.1 Entities. The images $\{I_1, \dots, I_N\}$ appearing in both Visual Genome [22] and train split dataset of MSCOCO or Flickr are selected, which can avoid the data leakage of the validation set. Then N ($N = 47210$) triple tuples (I_i, O_i, T_i) are gotten, where I_i is the raw image, O_i is the image object list appearing in the I_i , and T_i is the text caption annotated by human. We follow [18, 46] to ignore the meaningless textual words such as "is" and "a", and the n^t most frequently appearing text words $\{T'_1, T'_2, \dots, T'_{n^t}\}$ are selected from total 14777 textual words in $\{T_1, \dots, T_N\}$ and n^i most frequently appearing image objects $\{O'_1, O'_2, \dots, O'_{n^i}\}$ from total 56355 image objects in $\{O_1, \dots, O_N\}$.

3.2.2 Relations. The co-occurrence times between $\{T'_1, T'_2, \dots, T'_{n^t}, O'_1, O'_2, \dots, O'_{n^i}\}$ are counted according to the N triple tuples (I_i, O_i, T_i) . The co-occurrence matrix denotes $\mathbf{A} \in \mathbb{R}^{(n^t+n^i) \times (n^t+n^i)}$. WordNet's path similarity $s_p(\cdot, \cdot)$ [4, 32] is used to represent the intra-modal semantic relations. The path similarity matrix of text words denotes $\mathbf{A}^t \in \mathbb{R}^{n^t \times n^t}$ and the path similarity matrix of image objects denotes $\mathbf{A}^i \in \mathbb{R}^{n^i \times n^i}$. The two path similarity matrices are as follows:

$$\begin{cases} \mathbf{A}_{i,j}^t = s_p(T'_i, T'_j) \\ \mathbf{A}_{i,j}^i = s_p(O'_i, O'_j) \end{cases} \quad (1)$$

where $s_p(\cdot, \cdot)$ is calculated by "path similarity" in the Natural Language ToolKit (NLTK) [4]. The path similarity $s_p(\cdot, \cdot) = 1/(1 + d(\cdot, \cdot))$, where $d(\cdot, \cdot)$ represents the shortest path distance of two words in the is-a (hypernym/hyponym) taxonomy.

It returns a score denoting how similar two words are. The score is in the range 0 to 1, where 1 represents the maximum similarity and 0 represents the minimal similarity. The path similarity can help to distinguish other entities in the semantic space.

3.2.3 Representation of Entities. We follow CVSE [46] to embed each text entity T'_i into a vector $\mathbf{g}_i \in \mathbb{R}^{300}$ by Glove [35] rather than Bert [45]. It has two benefits: (1) The proposed approach MKVSE can be fairly compared with CVSE[†], which is a re-implementation of CVSE [46] using Bert as the input text's encoder and Glove as the text entities' encoder with slightly better results (See more details in section 4.4). (2) For the text encoder, Bert is usually better than Glove in performance [2]. The text entities in MKG are represented by Glove embedding and improve the final performance of image-text retrieval, which further shows the effectiveness of our proposed approach.

For each image object (entity) O'_i , raw images having shown the image object O'_i denotes $\{I_{i_1}, \dots, I_{i_{k_i}}\}$, where k_i is the total number of the above raw images. Each image object O'_i 's embedding $\mathbf{b}_i \in \mathbb{R}^{D_i}$ is represented as the average of all I_{i_j} 's R region-level image features $[\mathbf{v}_{j,1}, \dots, \mathbf{v}_{j,R}] \in \mathbb{R}^{R \times D_i}$ extracted from the Bottom-Up and Top-Down attention model (BUTD) [1] as follows:

$$\mathbf{b}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \frac{1}{R} \sum_{r=1}^R \mathbf{v}_{j,r} \quad (2)$$

The mean pooling of image features belonging to the same category are calculated, which can represent the average semantics of each object category.

Finally, the representations for text and image entities are acquired as:

$$\begin{cases} \mathbf{G} = [\mathbf{g}_1; \dots; \mathbf{g}_{n^t}] \\ \mathbf{B} = [\mathbf{b}_1; \dots; \mathbf{b}_{n^i}] \end{cases} \quad (3)$$

where n^i is the number of selected image objects most frequently appearing and n^t is the number of selected text words most frequently appearing.

Compared with the method that only utilize the word co-occurrences [41, 46], MKG has significant advantages. Obtaining the similarities of entities only based on word co-occurrences is easy to make mistakes. They are prone to conclude that "man" and "dog" are similar because these two words usually occur in the same image or the same sentences. However, MKG can address this problem to some extent. In fact, MKG can distinguish "man" and "dog" in the noun hierarchies of WordNet [32]. Moreover, the inter-modal co-occurrence relations characterize the co-occurrence correlations such as temporal, causal and logical relation which are implicit in the unimodal data (such as text corpus). For instance, the text entity "washing" and the image entity "tap", the text entity "cutting" and the image entity "knife" co-occur frequently in image-text pairs. Although the text entity does not directly describe the image entity, they are semantically related, which can be utilized to connect the two modalities.

3.3 Multimodal Graph Convolution Networks

Different modal entities would have different characteristics, and so as to different types of relations. To sufficiently utilize the implicit multimodal knowledge in MKG, including the two modal entities, intra-modal and inter-modal relations, Multimodal Graph Convolution Networks (MGCN) is proposed. MGCN reasons on MKG in two steps. In each step, MGCN can focus on different aspects of the implicit multimodal knowledge.

3.3.1 Intra-modal Relation Reasoning. In the first step, MGCN separately reason on the image entities and text entities connected by intra-modal semantic relations using Graph Convolution Networks (GCN) [21]. It can generate semantic features for each entity to distinguish other entities

in the semantic space. Concretely, the relation matrices \mathbf{A} , \mathbf{A}^t , \mathbf{A}^i are first normalized in Eqn. (1) as:

$$\begin{cases} \hat{\mathbf{A}}_{i,j} &= \frac{\mathbf{A}_{i,j}}{\sum_j \mathbf{A}_{i,j}} \\ \hat{\mathbf{A}}_{i,j}^i &= \frac{\mathbf{A}_{i,j}^i}{\sum_j \mathbf{A}_{i,j}^i} \\ \hat{\mathbf{A}}_{i,j}^t &= \frac{\mathbf{A}_{i,j}^t}{\sum_j \mathbf{A}_{i,j}^t} \end{cases} \quad (4)$$

MKG's entity embeddings \mathbf{G} and \mathbf{B} are projected into a D -dimensional space via a fully-connected (FC) linear projection to get $\mathbf{G}^{(0)} \in \mathbb{R}^{n^t \times D}$ and $\mathbf{B}^{(0)} \in \mathbb{R}^{n^i \times D}$. The GCN [21] is utilized to separately reason on text entities and image entities with intra-modal links as the first reasoning step:

$$\tilde{\mathbf{G}}^{(l)} = \begin{cases} \mathbf{G}^{(0)} & , l = 0 \\ \sigma \left(\hat{\mathbf{A}}^t \mathbf{G}^{(l-1)} \mathbf{W}_t^{(l-1)} + \mathbf{C}_t^{(l-1)} \right) & , 0 < l \leq l_m \end{cases} \quad (5)$$

$$\tilde{\mathbf{B}}^{(l)} = \begin{cases} \mathbf{B}^{(0)} & , l = 0 \\ \sigma \left(\hat{\mathbf{A}}^i \mathbf{B}^{(l-1)} \mathbf{W}_i^{(l-1)} + \mathbf{C}_i^{(l-1)} \right) & , 0 < l \leq l_m \end{cases} \quad (6)$$

where l_m is the total number of layers in MGCN, $\hat{\mathbf{A}}^t$ and $\hat{\mathbf{A}}^i$ are the adjacency matrices for MGCN during training, which is pre-calculated from MKG in Eqn. (4) and then fixed, $\mathbf{W}_t^{(l-1)} \in \mathbb{R}^{D \times D}$ and $\mathbf{W}_i^{(l-1)} \in \mathbb{R}^{D \times D}$ are learnable matrices, $\mathbf{C}_t^{(l-1)} \in \mathbb{R}^{n^t \times D}$ and $\mathbf{C}_i^{(l-1)} \in \mathbb{R}^{n^i \times D}$ are learnable bias matrices, and $\sigma(\cdot)$ is the LeakyReLU activation function.

3.3.2 Inter-modal Relation Reasoning. In the second step, the whole multimodal knowledge graph connected by inter-modal co-occurrence relations is reasoned on to generate representation for all entities in MKG. It can connect two modalities based on co-occurrence correlations $\hat{\mathbf{A}}$ in Eqn. (4). Then the inter-modal reasoning is implemented on the whole graph as follows:

$$\mathbf{M}^{(l)} = \sigma \left(\hat{\mathbf{A}} \left(\tilde{\mathbf{G}}^{(l)} \parallel \tilde{\mathbf{B}}^{(l)} \right) \mathbf{W}^{(l)} + \mathbf{C}^{(l)} \right) \quad (7)$$

$$\tilde{\mathbf{M}}^{(l)} = \ell_2 \left(\mathbf{M}^{(l)} + \left(\tilde{\mathbf{G}}^{(l)} \parallel \tilde{\mathbf{B}}^{(l)} \right) \right) \quad (8)$$

where ℓ_2 represents the ℓ_2 -norm function, \parallel represents the concatenating of two feature matrices along the feature dimension. And $\tilde{\mathbf{M}}^{(l)}$ will be divided into two modal entities as follows:

$$[\mathbf{G}^{(l)}; \mathbf{B}^{(l)}] = \tilde{\mathbf{M}}^{(l)} \quad (9)$$

where $\mathbf{G}^{(l)} \in \mathbb{R}^{n^t \times D}$ are the text entities' representation in the l th layer and $\mathbf{B}^{(l)} \in \mathbb{R}^{n^i \times D}$ are the image entities' representation in the l th layer. They will be used as input to the Eqn. (5, 6). Finally, our proposed MGCN(\cdot) can be formalized as:

$$\text{MGCN}(\mathbf{G}, \mathbf{B}, \hat{\mathbf{A}}, \hat{\mathbf{A}}^i, \hat{\mathbf{A}}^t) = \tilde{\mathbf{M}}^{(l_m)} \quad (10)$$

where l_m is the total number of layers in MGCN.

3.4 Embedding Enhancement

The input image's global embedding and the input text's global embedding are enhanced with representation of entities in MKG to generate the multimodal knowledge enhanced embeddings for similarity calculation. Specifically, the multihead attention mechanism [45] is adopted to encode the input image's global embedding $\tilde{\mathbf{v}}$ and the input text's global embedding $\tilde{\mathbf{t}}$ using MKG's entity embeddings $\tilde{\mathbf{M}}^{(l_m)}$ as follows:

$$\text{MultiHead}(\mathbf{X}, \mathbf{Y}) = \text{Concat}(\mathbf{h}_1, \dots, \mathbf{h}_H) + \mathbf{X}, \quad (11)$$

1:10

Feng et al.

where $\mathbf{X} = \bar{\mathbf{v}}$ or $\bar{\mathbf{t}}$, $\mathbf{Y} = \tilde{\mathbf{M}}^{(l_m)}$, $\text{Concat}(\cdot)$ represents the concatenation operation along the feature dimension, H denotes the number of heads, and the scaled-dot product attention $\text{Att}(\cdot)$ is used to calculate \mathbf{h}_i as follows:

$$\mathbf{h}_i = \text{Att}(\mathbf{X}\mathbf{W}_i^Q, \mathbf{Y}\mathbf{W}_i^K, \mathbf{Y}\mathbf{W}_i^V) \quad (12)$$

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad (13)$$

where softmax function is operated on each row and d_k the channel number of \mathbf{Q} and \mathbf{K} , and \mathbf{W}_i^Q , \mathbf{W}_i^K and \mathbf{W}_i^V are learnable matrices. Then, a fully connected feed-forward network is executed to combine attention results from different heads.

The global embeddings of two modalities are enhanced with the multi-head attention, which will be aligned in the semantic space formed by key values (entity embeddings generated by MGCN). The embedding enhancement can help injecting the multimodal knowledge relations into the final output embeddings. Because MKG can represent implicit relations between text and image modalities which are not contained in the global embedding, this enhancement can help the model to learn the implicit connection between text and image for a better image-text retrieval performance.

Based on the above processes, the multimodal knowledge enhanced embeddings can be used as follows:

$$\begin{cases} \mathbf{v}_a = \text{FFN}(\text{MultiHead}(\bar{\mathbf{v}}, \tilde{\mathbf{M}}^{(l_m)})) \\ \mathbf{t}_a = \text{FFN}(\text{MultiHead}(\bar{\mathbf{t}}, \tilde{\mathbf{M}}^{(l_m)})) \end{cases} \quad (14)$$

Where $\text{FFN}(\cdot)$ denotes the feed forward network implemented by a two-layer multi-layer perceptron with the ReLU activation function in between. The final output embeddings can be formulated as follows:

$$\begin{cases} \mathbf{v}_e = [\sqrt{1-\lambda_c}\bar{\mathbf{v}}, \sqrt{\lambda_c}\mathbf{v}_a], \\ \mathbf{t}_e = [\sqrt{1-\lambda_c}\bar{\mathbf{t}}, \sqrt{\lambda_c}\mathbf{t}_a], \end{cases} \quad (15)$$

where λ_c is the hyper-parameter of concatenating weight for the multimodal knowledge enhanced embeddings. The reason of using $\sqrt{\lambda_c}$ and $\sqrt{1-\lambda_c}$ as the concatenating weight will be explained in Section 3.5. The benefit of concatenating two parts of embeddings is efficient during inference for using embedding based retrieval.

3.5 Objective Function

To achieve alignment of a given positive image-text pair (I, T) , the hinge-based bidirectional triplet loss [14] is utilized for optimization, which is defined as:

$$L = [\alpha - s(I, T) + s(I, \hat{T})]_+ + [\alpha - s(I, T) + s(\hat{I}, T)]_+ \quad (16)$$

where α represents the margin factor, $[x]_+ = \max(x, 0)$, and $s(I, T)$ denotes the cosine similarity between the output embeddings of I and T (\mathbf{v}_e and \mathbf{t}_e). $\hat{T} = \arg\max_{j \neq T} s(I, j)$ and $\hat{I} = \arg\max_{i \neq I} s(i, T)$ are the hardest negatives in a mini-batch. $s(I, T)$ can be formulated as follows:

$$\begin{aligned} s(I, T) &:= \cos(\mathbf{v}_e, \mathbf{t}_e) \\ &= \cos([\sqrt{1-\lambda_c}\bar{\mathbf{v}}, \sqrt{\lambda_c}\mathbf{v}_a], [\sqrt{1-\lambda_c}\bar{\mathbf{t}}, \sqrt{\lambda_c}\mathbf{t}_a]) \\ &= \cos(\sqrt{1-\lambda_c}\bar{\mathbf{v}}, \sqrt{1-\lambda_c}\bar{\mathbf{t}}) + \cos(\sqrt{\lambda_c}\mathbf{v}_a, \sqrt{\lambda_c}\mathbf{t}_a) \\ &= (1-\lambda_c) \cos(\bar{\mathbf{v}}, \bar{\mathbf{t}}) + \lambda_c \cos(\mathbf{v}_a, \mathbf{t}_a) \end{aligned} \quad (17)$$

where $\|\bar{\mathbf{v}}\|_2 = \|\bar{\mathbf{t}}\|_2 = \|\mathbf{v}_a\|_2 = \|\mathbf{t}_a\|_2 = 1$. $\|\cdot\|_2$ represents the ℓ_2 -norm. $\cos(\cdot)$ represents the cosine similarity. The cosine similarity $s(I, T)$ has two parts: $(1-\lambda_c) \cos(\bar{\mathbf{v}}, \bar{\mathbf{t}}) = (1-\lambda_c)s(\bar{\mathbf{v}}, \bar{\mathbf{t}})$ and

$\lambda_c \cos(\mathbf{v}_a, \mathbf{t}_a) = \lambda_c s(\mathbf{v}_a, \mathbf{t}_a)$, where $s(\bar{\mathbf{v}}, \bar{\mathbf{t}})$ represents the cosine similarity of the input image's and the input text's global embedding, $s(\mathbf{v}_a, \mathbf{t}_a)$ represents the cosine similarity of the image's and text's multimodal knowledge enhanced embedding. λ_c is the weight of $s(\mathbf{v}_a, \mathbf{t}_a)$. The top similarity will be selected as the final retrieval result and the triplet loss is used for alignment learning.

4 EXPERIMENTS

To demonstrate the effectiveness of the proposed MKVSE approach, we perform experiments in terms of image-to-text retrieval (image query) and text-to-image retrieval (text query) on two widely-used datasets and compare with recent state-of-the-art methods. Ablation studies are conducted to investigate the effectiveness of each component of our approach. We also introduce detailed implementations and training strategy of the proposed MKVSE approach.

4.1 Datasets

Experiments on the two widely-used datasets: Flickr30K [56] and MSCOCO [28] are conducted to evaluate our approach and recent state-of-the-art methods. The details of the two datasets are as follows:

Flickr30K. It contains 31,783 images from Flickr website, and each image is described by 5 different sentences. Following the settings in [7, 37, 46], this dataset is split into 29,783 training images, 1,000 validation images, and 1,000 testing images.

MSCOCO. It includes 123,287 images, where each image is associated with 5 annotated sentences. Similarly, we followed the split of [7, 37, 46], namely 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. Likewise, two evaluation settings are considered in this paper: (1) MSCOCO 1K, the final result is calculated by averaging the results over 5-folds of 1K test images; and (2) MSCOCO 5K, the evaluation result is calculated on the full 5K testing images.

4.2 Experimental Settings

4.2.1 Evaluation Protocols. Following the existing methods [7, 37], we adopt Recall at K, R@K (K=1, 5, and 10) for short, as the evaluation metrics, which are commonly utilized in the multimodal retrieval task. R@K is defined as the percentage of ground truth being retrieved at top-K results. The higher R@K indicates the better performance. We also adopt RSUM (sum of R@K) as the evaluation metrics, which calculates the total value of R@K for both text and image retrieval. RSUM provides general perspective for the overall retrieval performance. Same as R@K, the higher RSUM indicates the better performance.

4.2.2 Implementation Details. Our implementation settings follow our baseline model GPO's [7]. For each image, the Faster-RCNN [38] detector provided by Bottom-Up and Top-Down attention model (BUTD) [1] are taken to extract the R ($R = 36$) region proposals and obtain a 2048-dimensional feature for each region. And the BUTD model is pre-trained on ImageNet [39] and Visual Genome [22] datasets. For each input text, The basic version of the pre-trained Bert [45] is leveraged to obtain the original word embeddings with dimension 768. The number of text entities and image entities are such that $n^t = n^i = 300$. Then they are projected to a D -dimensional space ($D = 1024$). The activation function $\sigma(\cdot)$ in this paper represents LeakyReLU function and its negative slope is 0.1. H in Eqn. (11) is 1. The concatenating weight λ_c in Eqn. (15) is 0.05. The total number of layers l_m in MGCN is 1. The model is trained with the batch size of 128. The margin α of the triplet ranking loss in Eqn. (16) is 0.2. The initial learning rate is $5e-4$ while different approach components have different learning rate multiplier: (1) Bert: 0.1; (2) MGCN: 0.5. The approach is trained for 25 epochs and learning rate decays by a factor of 10 for last 10 epochs. All experiments are implemented with PyTorch v1.2.0 and run on GTX 1080 Ti.

Table 1. Image-text retrieval performance on Flickr30K test set.

Methods	Backbone	Flickr30K Dataset						
		Image-to-Text			Text-to-Image			RSUM
		R@1	R@5	R@10	R@1	R@5	R@10	
SCAN* (ECCV 2018) [23]	BUTD, GRU	67.4	90.3	95.8	48.6	77.7	85.2	465.0
BFAN* (MM 2019) [29]	BUTD, GRU	68.1	91.4	-	50.8	78.4	-	-
VSRN* (ICCV 2019) [24]	BUTD, GRU	71.3	90.6	96.0	54.7	81.8	88.2	482.6
CVSE (ECCV 2020) [46]	BUTD, GRU	73.5	92.1	95.8	52.9	80.4	87.8	482.5
IMRAM (CVPR 2020) [6]	BUTD, GRU	74.1	93.0	96.6	53.9	79.4	87.2	484.2
WCGI (ICCV 2021) [51]	BUTD, GRU	74.8	93.3	96.8	54.8	80.6	87.5	487.8
ADAPT* (AAAI 2020) [52]	BUTD, GRU	76.6	95.4	97.6	60.7	86.6	92.0	508.9
SGRAF (AAAI 2021) [10]	BUTD, GRU	77.8	94.1	94.1	97.4	58.5	83.0	504.9
CAMERA* (MM 2020) [36]	BUTD, Bert	78.0	95.1	97.9	60.3	85.9	91.7	508.9
GraDual* (WACV 2022) [30]	MNET, GRU	78.3	<u>96.0</u>	98.0	60.4	86.7	92.0	511.4
VSRN++ (TPAMI 2022) [24]	BUTD, Bert	79.2	94.6	97.5	60.6	85.6	91.4	508.9
DIME* (SIGIR 2021) [37]	BUTD, Bert	81.0	95.9	<u>98.4</u>	<u>63.6</u>	<u>88.1</u>	<u>93.0</u>	<u>520.0</u>
GPO (CVPR 2021) [7]	BUTD, Bert	<u>81.7</u>	95.4	97.6	61.4	85.9	91.5	513.5
MKG (ours)	BUTD, Bert	80.1	95.8	98.6	63.2	87.4	92.3	517.4
MGCN (ours)	BUTD, Bert	82.9	96.5	98.9	63.2	87.1	92.5	521.1
MKVSE* (ours)	BUTD, Bert	84.0	96.9	99.1	64.4	88.2	93.1	525.7

Table 2. Image-text retrieval performance on MSCOCO 1K test set.

Methods	Backbone	MSCOCO (1K) Dataset						
		Image-to-Text			Text-to-Image			RSUM
		R@1	R@5	R@10	R@1	R@5	R@10	
SCAN* (ECCV 2018) [23]	BUTD, GRU	72.7	94.8	98.4	58.8	88.4	94.8	507.9
CVSE (ECCV 2020) [46]	BUTD, GRU	74.8	95.1	98.3	59.9	89.4	95.2	512.7
BFAN* (MM 2019) [29]	BUTD, GRU	74.9	95.2	-	59.4	88.4	-	-
WCGI (ICCV 2021) [51]	BUTD, GRU	75.4	95.5	98.6	60.8	89.3	95.3	514.9
VSRN* (ICCV 2019) [24]	BUTD, GRU	76.2	94.8	98.2	62.8	89.7	95.1	516.8
ADAPT* (AAAI 2020) [52]	BUTD, GRU	76.5	95.6	98.9	62.2	90.5	96.0	519.7
IMRAM (CVPR 2020) [6]	BUTD, GRU	76.7	95.6	98.5	61.7	89.1	95.0	516.6
GraDual* (WACV 2022) [30]	MNET, GRU	77.0	96.4	98.6	<u>65.3</u>	<u>91.9</u>	96.4	525.6
CAMERA* (MM 2020) [36]	BUTD, Bert	77.5	96.3	98.8	63.4	90.9	95.8	522.7
VSRN++ (TPAMI 2022) [24]	BUTD, Bert	77.9	96.0	98.5	64.1	91.0	96.1	523.6
DIME* (SIGIR 2021) [37]	BUTD, Bert	78.8	96.3	98.7	64.8	91.5	<u>96.5</u>	526.6
SGRAF (AAAI 2021) [10]	BUTD, GRU	79.6	96.2	98.5	63.2	90.7	96.1	524.3
GPO (CVPR 2021) [7]	BUTD, Bert	<u>79.7</u>	<u>96.4</u>	<u>98.9</u>	64.8	91.4	96.3	<u>527.5</u>
MKG (ours)	BUTD, Bert	79.8	96.7	98.8	65.6	91.5	96.2	528.6
MGCN (ours)	BUTD, Bert	79.8	96.5	98.9	65.0	91.6	96.4	528.2
MKVSE* (ours)	BUTD, Bert	81.0	96.5	99.0	66.4	92.1	96.6	531.6

4.3 Comparison with State-of-the-art Methods

To demonstrate the effectiveness of our proposed MKVSE approach, we compare it with the recent state-of-the-art methods in image-text retrieval task on two widely-used datasets. The comparison results are summarized in Table 1, Table 2, and Table 3. The best performance is highlighted in bold,

Table 3. Image-text retrieval performance on MSCOCO 5K test set.

Methods	Backbone	MSCOCO (5K) Dataset						
		Image-to-Text			Text-to-Image			RSUM
		R@1	R@5	R@10	R@1	R@5	R@10	
SCAN* (ECCV 2018) [23]	BUTD, GRU	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSRN* (ICCV 2019) [24]	BUTD, GRU	53.0	81.1	89.4	40.5	70.6	81.1	415.7
IMRAM (CVPR 2020) [6]	BUTD, GRU	53.7	83.2	91.0	39.6	69.1	79.8	416.4
VSRN++ (TPAMI 2022) [24]	BUTD, Bert	54.7	82.9	90.9	42.0	72.2	82.7	425.4
CAMERA* (MM 2020) [36]	BUTD, Bert	55.1	82.9	91.2	40.5	71.7	82.5	423.9
SGRAF (AAAI 2021) [10]	BUTD, GRU	57.8	-	91.6	41.9	-	81.3	-
GPO (CVPR 2021) [7]	BUTD, Bert	58.3	85.3	<u>92.3</u>	42.4	72.7	<u>83.2</u>	434.2
DIME* (SIGIR 2021) [37]	BUTD, Bert	<u>59.3</u>	<u>85.4</u>	91.9	<u>43.1</u>	<u>73.0</u>	83.1	<u>435.8</u>
MKG (ours)	BUTD, Bert	59.1	85.6	92.7	43.3	73.2	83.4	437.3
MGCN (ours)	BUTD, Bert	59.3	84.9	92.6	42.8	73.2	83.4	436.2
MKVSE* (ours)	BUTD, Bert	60.8	86.6	93.1	44.3	74.1	84.3	443.2

and the best performance of previous methods is emphasized with underlines. The state-of-the-art methods include SCAN [23], BFAN [29], VSRN [24], CVSE [46], IMRAM [6], WCGL [51], ADAPT [52], SGRAF [10], CAMERA [36], GraDual [30], VSRN++ [25], DIME [37], and GPO [7]. "MKG" in the tables represents only using MKG in our approach with Consensus-aware Graph Convolutional Network (CGCN) [46]. "MGCN" in the tables utilizes not only MKG but also MGCN to reason on MKG for getting the representations of MKG. Since some of state-of-the-art methods are ensemble models (marked with the symbol "*" in the Table), we follow [23, 37] to provide the ensemble model MKVSE* for fair comparison, which use averaging similarity scores of MKG and MGCN models for final evaluation. Quantitative results on Flickr30K are shown in Table 1. And the same retrieval task results on MS-COCO 1K test set and 5K test set are shown in Table 2 and Table 3 respectively. The column of "Backbone" represents the backbone of the corresponding method:

- "BUTD" represents the Bottom-Up and Top-Down attention model [1] for image encoding. This model builds on Faster-RCNN [38] and pre-trains on Visual Genome [22].
- "MNET" represents the Motif-Net [57] for image encoding. Motif-Net builds on Faster-RCNN [38] for predicting bounding regions, fine tuned and adapted for Visual Genome [22].
- "GRU" represents the Gate Recurrent Unit [3, 40] for text encoding.
- "Bert" represents the Bert model [45] for text encoding.

It is worth noting that, in the Table 1, Table 2, and Table 3, all state-of-the-art methods use the BUTD or Motif-Net as the image encoder, which are both pre-trained on Visual Genome dataset. And our approach also utilizes this dataset. In this paper, we follow the settings in [7, 36, 37] to use the same backbone, Bottom-Up and Top-Down attention model (BUTD) [1] and Bert [45] extracting features from images and texts.

Comparison of single models. Our proposed single model MGCN outperforms other state-of-the-art single methods on both two datasets in R@1, which is the hardest metric. Moreover, compared with the best performance of previous single models, our MGCN obtains relative RSUM gains with 7.6% on Flickr30K dataset in Table 1. Our MGCN also obtains relative RSUM gains with 0.7% and 2.0% on MSCOCO (1K) settings and MSCOCO (5K) settings in Table 2 and Table 3.

Comparison of ensemble models. As showing in Table 1, Table 2 and Table 3, our ensemble model MKVSE* outperforms other ensemble methods in all metrics, i.e. R@K (K = 1, 5, 10) and RSUM. Specifically, compared with the best performance of previous methods in image-to-text

retrieval and text-to-image retrieval tasks, MKVSE* gains 2.3% and 0.8% R@1 on Flickr30K, 1.3% and 1.6% R@1 on MSCOCO(1K), 1.5% and 1.2% R@1 on MSCOCO(5K). And MKVSE* also obtains relative RSUM gains with 5.7%, 3.1%, 7.4% on Flickr30K, MSCOCO (1K), and MSCOCO (5K) .

Analysis of the results. Table 1, Table 2 and Table 3 shows that the performances of all methods on the three settings (Flickr30K, MSCOCO (1K) and MSCOCO (5K)) are relatively consistent. In the following analysis, we focus on the performance of Flickr30K in Table 1. We can see that the performance of SCAN* and BFAN* is relatively low, because the simple architecture cannot fully capture the semantic and interaction of two modalities: only utilizing attention mechanism for connecting words and image regions. The follow-up works are mainly improved in two aspects: better similarity representation and better feature extractors. As for better similarity representation, DIME* utilizes a modality interaction modeling network based upon the routing mechanism to dynamically learn different activated paths for different data. As for better feature extractors, GraDual* utilizes visual and textual scene graph and improves the coverage of each modality by exploiting textual context semantics for the image representation, and using visual features as a guidance for the text representation. Moreover, GPO utilizes a generalized pooling operator to automatically seek the best pooling function for different data modalities and feature extractors.

However, these methods only rely on the image-text pairs or utilize the additional unimodal knowledge (e.g. textual knowledge). They neglect the implicit multimodal knowledge relations between the image and text. When the image includes the object that is not directly described in the text, these methods can not connect images and texts well for lacking the guides of multimodal semantic relations. Our proposed approach introduces MKG to explicitly represent implicit multimodal knowledge relations between the image and text. And these relations can be sufficiently utilized by our proposed MGCN with a two-step reasoning. The benefits of our approach are as follows:

- **Multimodal Knowledge Graph.** MKG can explicitly represent implicit multimodal knowledge relations between the image and text, including intra-modal semantic relations and inter-modal co-occurrence relations. The intra-modal semantic relations can help to distinguish other entities in the semantic space and the inter-modal co-occurrence relations can connect two modal entities based on co-occurrence correlations. The image entities and the text entities may be not directly related in some situations, but the co-occurrence relations can depict their implicit correlations (Table 1, Table 2, and Table 3).
- **Multimodal Graph Convolution Network.** MGCN can sufficiently utilize the implicit multimodal knowledge by reasoning on the MKG in two steps, in each of which MGCN can focus on different aspects: the intra-modal reasoning to distinguish other entities in the semantic space and the inter-modal reasoning to connect multimodal entities based on co-occurrence correlations. The two-step reasoning manner can help to sufficiently mine the implicit semantic relations between two modal entities to enhance the representation of the image and text (Table 1, Table 2, and Table 3).

The best results over previous methods on all four metrics indicate the effectiveness and importance of implicit multimodal knowledge utilizing and injecting. Thus, the implicit multimodal knowledge can further strengthen the visual-semantic embedding.

4.4 Ablation Study

In this section, we conduct several experiments using the single model MKG and MGCN to further analyze the effectiveness of our approach. Specifically, we explore how each component of our approach, including the MKG and MGCN, affects the image-text retrieval results on Flickr30K (Table 4).

Table 4. Effectiveness of each component in our MKVSE approach.

Methods	Flickr30K Dataset						
	Image-to-Text			Text-to-Image			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
GPO	81.7	95.4	97.6	61.4	85.9	91.5	513.5
CVSE [†]	81.0	96.4	98.3	61.9	87.1	92.0	516.7
MKG (ours)	80.1	95.8	98.6	63.2	87.4	92.3	517.4
MGCN (ours)	82.9	96.5	98.9	63.2	87.1	92.5	521.1

4.4.1 Multimodal Knowledge Graph. In Table 4, GPO [7] is the baseline model. To demonstrate the effectiveness of MKG's introduce, we re-implement CVSE [46] with more powerful backbone Bert [45] as the encoder of the input text. And CVSE[†] is the re-implementation model with slightly better results than original CVSE's (see Table 1). CVSE[†] uses the textual knowledge compared with baseline GPO and achieves better results, which shows that introducing the textual knowledge can strengthen the semantic relations between image and text. Then we propose MKG, which explicitly represent the implicit multimodal knowledge relations between the image and text as intra-modal semantic relations and inter-modal co-occurrence relations. Intra-modal semantic relations provide synonymy information and inter-modal co-occurrence relations characterize the co-occurrence correlations. Our proposed MKG uses the addition multimodal knowledge graph and graph reasoning method CGCN which is the same graph reasoning method as CVSE[†]. MKG improves the RSUM as compared with the results of CVSE[†], which shows that introducing implicit multimodal knowledge relations can help to connect the image and text better. However, MKG performs less than CVSE[†] in the ablation experiments on image-to-text retrieval. This is because the previous graph reasoning method CGCN used by above two networks can not sufficiently utilize the multimodal knowledge in MKG. CGCN is proposed to utilize the unimodal knowledge.

4.4.2 Multimodal Graph Convolution Networks. MGCN is proposed to solve the problem above by decomposing the multimodal graph convolution networks into intra-modal relation reasoning and inter-modal relation reasoning. The benefit is that MGCN can focus on different aspects of multimodal knowledge in each step to sufficiently utilize the in Multimodal Knowledge Graph. In each step, MGCN can focus on different aspects of multimodal knowledge. In the first step, MGCN can focus on the intra-modal relations to distinguish other entities in the semantic space. In the second step, MGCN can focus on the intra-modal relations to connect multimodal entities based on co-occurrence correlations. The two-step reasoning manner can take the distinct characteristics between text entities and image entities into consideration, which helps to improve the quality of visual-semantic embedding. The comparison results on Flickr30K between MKG and MGCN are in Table 4. We can see distinct improvements after utilizing MGCN to extract features in MKG. As compared with MKG, MGCN gains 3.7% RSUM. As compared with CVSE[†], MGCN gains 4.4% RSUM and performs better than CVSE[†] on both image-to-text and image-to-text retrieval tasks, which demonstrates the effectiveness of the two-step reasoning of MGCN.

4.5 Parameter Analysis

To explore the impact of the hyper-parameter introduced by our approach, we conduct the parameter experiments to further analyze our approach, including the number of image entities n^i (Table 5), the concatenating weight for the multimodal knowledge enhanced embedding λ_c (Table 6) and the number of MGCN's layers l_m (Table 7). All ablation experimental results are conducted on Flickr30K.

Table 5. Effects of the number of image entities n^i on Flickr30K dataset.

n^i	Flickr30K Dataset						
	Image-to-Text			Text-to-Image			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
0	81.7	95.4	97.6	61.4	85.9	91.5	513.5
100	80.9	96.1	98.3	62.1	87.0	92.1	516.5
200	79.2	94.7	97.4	59.8	85.3	91.1	507.5
300	82.9	96.5	98.9	63.2	87.1	92.5	521.1
400	80.3	95.0	98.3	62.6	87.2	92.3	515.7

Table 6. Effects of the concatenating weight λ_c on Flickr30K dataset.

λ_c	Flickr30K Dataset						
	Image-to-Text			Text-to-Image			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
0	81.7	95.4	97.6	61.4	85.9	91.5	513.5
0.025	78.6	93.7	97.4	59.8	85.4	90.9	505.8
0.05	82.9	96.5	98.9	63.2	87.1	92.5	521.1
0.075	81.0	96.0	98.1	63.5	87.6	92.7	518.9
0.1	27.7	58.4	73.0	19.7	48.8	64.7	292.3

Table 7. Effects of the number of MGCN layers l_m on Flickr30K dataset.

l_m	Flickr30K Dataset						
	Image-to-Text			Text-to-Image			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
1	82.9	96.5	98.9	63.2	87.1	92.5	521.1
2	80.7	94.7	97.6	63.0	87.0	92.4	515.4
3	81.8	94.5	97.6	62.9	86.9	92.3	516.0

- To perform a sensitivity analysis of the parameters n^i , we conduct experiments by increasing it from 0 to 400. The results are shown in Table 5. It can be seen that increasing n^i doesn't always benefit for the performance. In fact, when n^i increasing from 300 to 400, the RSUM has a drop of 5.4 %. This can be caused by the latter 100 words appearing less frequently, which introduce the noise to the model to hinder the ability of learning robust concept features.
- To explore the impact of the parameters λ_c , it was fine-tuned and see how the performance of the models varies. In Table 6, it can be found that the model is a little sensitive to the λ_c . When λ_c is set to be 0.1, the RSUM of our model will drop to 292.3. However, the concatenating weight for concept embedding in proper range will enhance the retrieval performance such as 0.05 and 0.075.
- We change the l_m from 1 to 3, the performance is not better with the deeper MGCN. One possible reason is that GCNs are hard to be deep due to the over-smoothing and gradient vanishing problems [26].

























Text Query	MKVSE (ours)			GPO		
Fruits and vegetables are on the stove top in kitchen.						
A face of flowers is below an ornamental gold and white ceiling.						
The control stand used by the train conductor.						
A small plate with some vegetables being held by a person.						

Fig. 3. Visual comparisons of text-to-image retrieval examples between baseline GPO and our MKVSE on MSCOCO test dataset. The ground truth of images are outlined in green boxes and the incorrect ones are outlined in red boxes.





Image Query	MKVSE (ours)	GPO
	<ol style="list-style-type: none"> 1. A boy reaching for a toothbrush in a store from a buggy. 2. A young boy taking a toothbrush off the store shelf. 3. A little boy picking up a toothbrush in a store. 4. A boy is picking out toothbrushes at a store. 5. A young boy in a shopping cart looking at tooth brushes. 	<ol style="list-style-type: none"> 1. A young boy taking a toothbrush off the store shelf. 2. A boy reaching for a toothbrush in a store from a buggy. 3. A boy tries a bicycle in front of a display stand. 4. A boy is picking out toothbrushes at a store. 5. A little boy picking up a toothbrush in a store.
	<ol style="list-style-type: none"> 1. A large head of broccoli in a green garden. 2. Up close picture of broccoli with plants behind it. 3. A head of broccoli grows in a garden. 4. Broccoli is growing in the bright shiny sunlight. 5. A close up of some freshly grown broccoli. 	<ol style="list-style-type: none"> 1. A lot of plants , there tops green and stalks are brown. 2. Broccoli is growing in the bright shiny sunlight . 3. A large head of broccoli in a green garden. 4. A head of broccoli grows in a garden. 5. Up close picture of broccoli with plants behind it.
	<ol style="list-style-type: none"> 1. A young girl is trying to brush her hair with a pink brush. 2. A little girl with a bunny shirt brushing her hair with a pink brush. 3. A young child brushing her hair with a big pink brush. 4. A young girl tries to comb her own hair. 5. A young girl is combing her hair and looking at the camera. 	<ol style="list-style-type: none"> 1. A young girl is trying to brush her hair with a pink brush. 2. A little girl with a bunny shirt brushing her hair with a pink brush. 3. A young child brushing her hair with a big pink brush. 4. A young girl tries to comb her own hair. 5. A little girl is brushing her hair in a bathroom.
	<ol style="list-style-type: none"> 1. A man milking a brown and white cow in bam . 2. A man milking a cow during the day . 3. A man on a stool milking a cow . 4. The guy with the white shirt and baseball cap is milking the cow. 5. A man sitting on a stool milking a cow. 	<ol style="list-style-type: none"> 1. A man milking a brown and white cow in bam. 2. A man milking a cow during the day. 3. A man spreading out some hay in a large animal pen. 4. The guy with the white shirt and baseball cap is milking the cow. 5. A man on a stool milking a cow.

Fig. 4. Visual comparisons of image-to-text retrieval examples between baseline GPO and our MKVSE on MSCOCO test dataset. The ground truth texts are green, and the incorrect texts are red.

4.6 Qualitative Results

To better understand the effectiveness of our proposed approach MKVSE, we visualize some examples of image-to-text retrieval and text-to-image retrieval on MSCOCO test split dataset. For

each text query shown in Figure 3, the top 3 ranked images for our approach MKVSE and baseline model GPO are listed. The ground truth of images are outlined in green boxes and the incorrect ones are outlined in red boxes. The blue words in the text query are the key text entities. In the first example, the retrieval results of baseline model GPO do not contain fruits which is in the text query. However, our approach MKVSE enhances the ground truth image with image entities "orange" and "apple", which benefits to connect the image to word "fruits". The implicit multimodal knowledge relations between image entities "orange" "apple" and text entity "fruits" help to retrieve the corresponding image here. In the second example, baseline method ranks wrongly in this query because the retrieval images don't contain "ceiling" in the text query. However, our MKVSE enhance the ground truth image with image entity "lighting", which usually appears with text entity "ceiling". This multimodal relations help to retrieve the images with the object "ceiling". The similar implicit multimodal relations among the text query and retrieval images in the third and fourth examples, such as text entity "train" and image entity "railway" can also be observed.

Figure 4 shows the retrieval texts of each image query, the top 5 ranked texts are listed for our approach MKVSE and baseline model GPO. The ground truth texts are green, and the incorrect texts are red. It can be observed that our approach MKVSE is more robust in complex scenes than baseline GPO, achieving promising retrieval results. Such as the first example, the baseline GPO neglects the implicit relations between "toothbrush" in the image and "bicycle" in the text, which do not co-occurrence often. In fact, "bicycle" often appearances outdoor (such as road) but "toothbrush" often appearances indoor (such as on the washbasin and the store shelf). Our MKVSE can correct this error in some content by utilizing implicit multimodal knowledge relations.

5 CONCLUSION

In this paper, we have proposed the **Multimodal Knowledge enhanced Visual-Semantic Embedding (MKVSE)** approach for image-text retrieval, which utilizes multimodal knowledge graph in embedding based image-text retrieval. Concretely, first the **Multimodal Knowledge Graph (MKG)** is built to represent the implicit multimodal knowledge relations between the image and text. Then **Multimodal Graph Convolution Networks (MGCN)** is introduced to reason on MKG in two steps for the visual-semantic embedding. Extensive experimental results on two benchmarks have demonstrated the effectiveness and superiority of our proposed method.

The future work lies in mainly two aspects. Firstly, it is important to utilize more types of multimodal knowledge relations between different modalities in multimodal retrieval systems, for improving the performance and explainability of the multimodal analysis. Secondly, more modalities will be incorporated for extending the field of multimodal knowledge relations, such as audio and video.

REFERENCES

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.
- [2] Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual Embeddings: When Are They Worth It?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 2650–2663. <https://doi.org/10.18653/v1/2020.acl-main.236>
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [4] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- [5] Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021. Multimodal named entity recognition with image attributes and image knowledge. In *International Conference on Database Systems for Advanced Applications*. Springer, 186–201.

- [6] 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/CVF conference on computer vision and pattern recognition*. 12655–12663.
- [7] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15789–15798.
- [8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.
- [9] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. 2022. Cross-modal Graph Matching Network for Image-text Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 4 (2022), 1–23.
- [10] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1218–1226.
- [11] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. 2022. MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5089–5098.
- [12] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 467–474.
- [13] 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*. 4601–4611.
- [14] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).
- [15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems* 26 (2013).
- [16] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. 2019. Context-aware embeddings for automatic art analysis. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 25–33.
- [17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.
- [18] Jingyi Hou, Xinxiao Wu, Wentian Zhao, Jiebo Luo, and Yunde Jia. 2019. Joint syntax representation learning and visual cue translation for video captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8918–8927.
- [19] Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. 639–645.
- [20] 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*. 3128–3137.
- [21] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=SJU4ayYgl>
- [22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.
- [23] 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 (ECCV)*. 201–216.
- [24] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4654–4662.
- [25] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2022. Image-Text Embedding Learning via Visual and Textual Semantic Reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [26] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI conference on artificial intelligence*.
- [27] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer,

1:20

Feng et al.

740–755.

- [29] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus Your Attention: A Bidirectional Focal Attention Network for Image-Text Matching. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 3–11. <https://doi.org/10.1145/3343031.3350869>
- [30] Siqu Long, Soyeon Caren Han, Xiaojun Wan, and Josiah Poon. 2022. Gradual: Graph-based dual-modal representation for image-text matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3459–3468.
- [31] Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 164–172.
- [32] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
- [33] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 299–307.
- [34] Fudong Nian, Bing-Kun Bao, Teng Li, and Changsheng Xu. 2017. Multi-modal knowledge representation learning via webly-supervised relationships mining. In *Proceedings of the 25th ACM international conference on Multimedia*. 411–419.
- [35] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
- [36] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. 2020. Context-aware multi-view summarization network for image-text matching. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1047–1055.
- [37] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1104–1113.
- [38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015), 91–99.
- [39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.
- [40] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.
- [41] Botian Shi, Lei Ji, Pan Lu, Zhendong Niu, and Nan Duan. 2019. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. In *IJCAI*, Vol. 1. 2.
- [42] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.
- [43] Yale Song and Mohammad Soleymani. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1979–1988.
- [44] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1405–1414.
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [46] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020. Consensus-aware visual-semantic embedding for image-text matching. In *Proceedings of the European Conference on Computer Vision*. Springer, 18–34.
- [47] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 394–407.
- [48] 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/CVF winter conference on applications of computer vision*. 1508–1517.
- [49] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 540–547.
- [50] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position Focused Attention Network for Image-Text Matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. 3792–3798.

- [51] Yun Wang, Tong Zhang, Xueya Zhang, Zhen Cui, Yuge Huang, Pengcheng Shen, Shaoxin Li, and Jian Yang. 2021. Wasserstein Coupled Graph Learning for Cross-Modal Retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 1793–1802.
- [52] 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*, Vol. 34. 12313–12320.
- [53] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-Language Pre-Training with Triple Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15671–15680.
- [54] Shiquan Yang, Rui Zhang, Sarah M Erfani, and Jey Han Lau. 2021. UniMF: A Unified Framework to Incorporate Multimodal Knowledge Bases into End-to-End Task-Oriented Dialogue Systems.. In *IJCAI*. 3978–3984.
- [55] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3025–3035.
- [56] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.
- [57] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5831–5840.
- [58] Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020. Poet: Product-oriented video captioner for e-commerce. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1292–1301.
- [59] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52, 1 (2019), 1–38.

Received August 2022; revised December 2022; accepted January 2023