



Flexible graph-based attention and pooling network for image-text retrieval

Hao Sun¹ · Xiaolin Qin¹ · Xiaojing Liu¹

Received: 24 September 2023 / Revised: 15 November 2023 / Accepted: 30 November 2023 /

Published online: 16 December 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

To explore the cross-modal semantic relevance, image-text retrieval has attracted much attention from the research community. The most of prominent models have shown that introducing structured graph information is capable of improving the cross-modal retrieval performance. However, the existing models either only focus on exploiting the structured information within individual modality, or learn specific cross-modal graph-level metric functions which fail to construct a shared semantic subspace for efficient retrieval. In this paper, we propose a graph attention network to transform general structured visual and textual graphs into the shared semantic subspace. Specially, a structured semantic enhancement module is proposed to learn the graph-level relevance information between images and sentences, which is further utilized to promote the cross-modal semantic alignment. And the enhancement module only depends on the structured input information at retrieval stage, which endows our model with the flexibility that processing fragment-level data no matter whether the structured information lacks. Besides, a graph-based pooling network is proposed to transform the fragment-level features to the common cross-modal representations for efficient retrieval. When compared with several state-of-the-art baselines, the experiments show that our model achieves competitive performance on two publicly available datasets Flickr30k and MS-COCO.

Keywords Common representation · Graph data · Attention mechanism · Image-text retrieval

Xiaolin Qin and Xiaojing Liu are both are equally contributed to this work

✉ Xiaolin Qin
qinxcs@nuaa.edu.cn

Hao Sun
sunhao123@nuaa.edu.cn

Xiaojing Liu
liuxiaojing@nuaa.edu.cn

¹ College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

1 Introduction

As the amount of available multi-modality data drastically increases, the demand for various intelligent multi-modality applications is imperative, e.g., indoor navigation [1, 2], video captioning [3], web information extraction [4–6] and so on. As a common task, image-text retrieval has attracted considerable attention from research community. In recent years, the most of representative methods tend to evaluate semantic relevance by exploring the similarity at the level of regions and words, which is termed as fragment-level matching in this paper. Some early works [7, 8] attempt to evaluate the similarity of image-sentence pairs by multi-step aggregation, in which the models selectively focus on different informative region-word pairs at each time step. The most of recent works [9–15] have focused on directly evaluating the image-sentence relevance given two sets of fragment-level features by introducing a variety of attention mechanisms. Lee et al. propose a basic cross-attention mechanism to compute the similarity between two feature sets of images and sentences [9]. Except for the similar attention mechanism, much effort has been made to further improve retrieval performance by introducing auxiliary topic constraint [10], negative-aware constraint [11] or global context reference [13]. Besides, the other works also attempt to integrate fragment-level semantic relevance with information flow controlling [16], action-aware memory retrieving [17] and multi-level feature fusion strategy [18, 19]. The aforementioned models have made great effort to promote cross-modal retrieval performance by exploring fine-grained fragment-level relevance. However, as reported in recent works [20], both attention and fusion mechanism require cross-modal operation over image and text representations, which make it infeasible to explicitly construct a shared cross-modal semantic subspace for efficient retrieval.

Instead of only exploiting fragment-level semantic features, some works [21–25] have attempted to introduce structured information to further promote the cross-modal semantic alignment. Because the visual and textual graph are common structured information in these works, we divide these graph-based methods into the category named graph-level matching in this paper. Li et al. propose to construct the visual graphs of images with structured information learned from region-based features [21], in which the visual graph is fully connected and undirected on account of little prior knowledge. The other works [22–25] not only incorporate specific auxiliary prior constraints to construct more reasonable and informative structured graphs, but also convert all sentences to textual semantic graphs. Then some multi-level feature fusion and attention-based strategies are utilized to perform graph-level image-sentence matching. The graph-based models have shown that structured semantic information is beneficial to the cross-modal retrieval. However, the most of graph-based models also fail to provide the common cross-modal representations, i.e., the feature vectors in a shared semantic subspace, and only exploit the structural information within individual modality.

In conclusion, both fragment-based and graph-based models have achieved considerable cross-modal retrieval performance. However they still face two-fold challenges which hinder the development of image-sentence retrieval. The one is that the most of them fail to construct a shared semantic subspace where images and sentences are represented as points and retrieval is equivalent to the vector-based ranking problem. The other one is that the most of graph-based models generally exploit the structural information to only refine intra-modal semantic features. To mitigate these problems, we propose a flexible graph attention network to perform graph-level cross-modal retrieval in this paper. To make full use of the structured information, a structured semantic enhancement module is proposed to learn the shared structural features for improving fragment-level representations and promoting cross-

modal semantic alignment. And the enhancement module only take as input the structured information, which endows our model with the capability of processing fragment-level data no matter whether the structured information lacks. To explicitly construct the shared semantic subspace, we propose a graph-based pooling module to project a set of fragment-level features to a feature vector in the common semantic subspace. The following experiments show that the proposed model can achieve competitive retrieval performance on two publicly available datasets when compared with several existing state-of-the-art models.

2 Related work

2.1 Fragment-level matching

To infer more accurate cross-modal semantic relevance, many existing methods attempt to explore the similarity at the level of regions and words, which is termed as fragment-level matching. Huang et al. propose to selectively focus on different informative region-word pairs given global context reference at multiple time steps and utilize LSTM [26] to aggregate all similarity vectors at various time steps [7]. Similarly, Nam et al. propose the dual attention network to compute fragment-level similarity scores at multiple time steps and retrieve database items with respect to the sum of all scores [8]. Lee et al. propose the stacked cross-modal attention network to aggregate fragment-level semantic relevance [9]. They first apply the attention mechanism to determine importance distribution over all fragments, and then compute the importance-based weighted sum of all fragments as the final feature vectors which is used to calculate similarity between holistic images and sentences. Follow the similar framework, some other methods [12, 27–30] attempt to improve the cross-modal attention-based retrieval by introducing either global context reference or prior knowledge-based constraints.

Except for focusing on inter-modal alignment, Wei et al. first enhance intra-modal semantic information with the self-attention mechanism and then refine inter-modal relevance with the cross-modal attention mechanism [15]. Yu et al. first enhance the fragment-level features with a multi-layer cross-modal attention module and then compute the similarity scores between pairwise enhanced features with a heterogeneous attention module [14]. Zhang et al. design a context-aware attention network to selectively focus on both intra-modal and inter-modal informative fragments given the global context [13]. Zhang et al. propose the negative-aware attention mechanism to take both matched and mismatched region-word pairs into consideration, and design a dynamic updating strategy to select positive and negative sample sets [11]. Wu et al. [10] propose the region reinforcement attention network to differentially attend to various region-word pairs while calculating semantic similarity. And they design a topic-based constraint module to promote the cross-modal semantic alignment. Qu et al. apply routing mechanism to dynamically control cross-modal information interaction to selectively aggregate fragment-level features with respect to the input samples [16]. Li et al. introduce the action-aware information to improve the common cross-modal representations [17], in which an action predictor is utilized to determine the action tags and the response representations in an action memory bank [17]. Lan et al. [19] propose a multi-level fusion matching strategy to integrate local and global features, in which the fusion representation is transformed to the similarity score with the multi-head attention mechanism and fully connected network [19].

The aforementioned methods have made great effort to explore and aggregate fine-grained semantic relevance, but fail to project both images and sentences into a shared feature sub-

space where the retrieval can be efficiently completed by distance-based ranking. Specially, there are also some fragment-based methods which are capable of providing common representations. Qu et al. propose the context-aware summarization network to match sentences with multi-view fused image representations [18], however, in which the max pooling operation used to compute similarity scores still brings significant computation burden when dealing with a huge amount of samples. Wu et al. propose to learn fragment-level embedding with multi-head self-attention mechanism [31] and apply the average pooling operation to generate the cross-modal representations of images and sentences [32]. Instead of average pooling operation, we propose a graph-based pooling module to embed structured information into the process of aggregating multiple fragment-level feature vectors.

2.2 Graph-level matching

The recent works attempt to introduce structural semantic information (e.g., connectivity between fragments) to construct multi-modal structured graphs for retrieval. Li et al. propose to construct fully connected visual graphs with respect to all salient region-based features [21], and then utilize the GCN [33] to further refine visual graphs. Finally, the enhanced visual graphs are transformed to the common cross-modal representations with a reasoning GRU [34]. Liu et al. propose a graph structured matching network to explore graph-level semantic correspondence between images and sentences [22], in which both node-level and structure-level similarity are taken into consideration. Except for learning the structured semantic information from scratch, some other works tend to construct more reasonable and informative graphs by incorporating auxiliary prior knowledge-based constraints and superior graph generators, e.g., visual scene graph generators [35, 36] trained on the datasets [37–39]. Wang et al. attempt to convert both images and sentences to semantic graphs and integrate the object-level and relationship-level similarity as the semantic relevance between holistic images and sentences [24]. Similarly, Zhong et al. construct the bi-level visual and textual graphs and compute both node-level and structure-level semantic similarity [23]. Lu et al. attempt to generate the hash code for complete or incomplete multimedia items with the multi-modal fusion graph. And a semantic GCN module is applied to supervise the hash learning of a hash GCN [40]. Ge et al. propose a structured multi-modal embedding network to learn robust cross-modal representations by aggregating instance-level [41], context-aware structured and consensus-aware concept [42] semantic features. Long et al. attempt to leverage cross-modal semantic cues to promote the construction of two uni-modal scene graphs [25]. In detail, the visual position information is introduced to generate the vision-integrated text embedding for each sentence. And the prior semantic knowledge is introduced to generate the context-integrated visual embedding for each image. A dual graph-based matching strategy is proposed to perform image and sentence retrieval independently. Besides, attention mechanism has also been widely applied in a variety of graph-based tasks. Yan et al. designed a hierarchical attention fusion mechanism for geo-location [43]. Cui et al. applied the attention-based blender module to combine the temporal relation and neighboring feature in video objection detection [44]. Liu et al. proposed a multi-scale feature aggregation strategy to selectively focus on key points in visual localization task [45]. Cui et al. proposed a geometric attentional edge convolution module to learn point cloud representations from both intrinsic and extrinsic properties [46]. Similar graph attention mechanism is also utilized to refine the visual and textual features in image-text retrieval. The most graph-based methods either fail to construct a explicit semantic subspace or only leverage the structured information to refine intra-modal fragment-level representations. In contrast, the proposed

model not only explicitly provides the common cross-modal representations for images and sentences, but also promote the cross-modal semantic alignment by learning the shared structured information between visual and textual graphs.

3 Flexible graph attention network

In this paper, we propose a flexible graph attention network to perform graph-level image-sentence retrieval. The Fig. 1 presents an overview of the proposed model. As illustrated in the Fig. 1, both images and sentences are first transformed to the data of type graph. And then our model projects both visual and textual graphs into a shared semantic subspace. In this work, we regard the graph data as the combination of fragment (vertex) and structured (edge or relationship) semantic information. The fragment and structured information flows are depicted as green and blue directed arrows in the Fig. 1. Besides, the fusion information flow is depicted as the brown directed arrows. Finally, we model the cross-modal semantic subspace with the rank loss and mutual information estimation loss. In this section, we will elaborate on the proposed model from three aspects, the generation of visual and textual graphs in the Section 3.1, the model architecture in the Section 3.2 and the training strategy in the Section 3.3.

3.1 Semantic graph generation

As in many previous works, we first apply the pretrained Faster RCNN [47] model to detect salient regions in each image. The 36 proposals with top confidence scores are selected as the visual fragment-level features. The averages of corresponding outputs in pooling layer are extracted as the feature vectors of proposals. Given these proposals and tentative classification results, we further build the visual scene graph using the Causal TDE [48] and NeuralMotifs [35] algorithms, in which the categories of all proposals are refined and all possible relationships are detected and classified. Finally, each image is represented as a graph $\mathcal{G}_V = \{V, C_v, C_r\}$, where $V \in R^{M \times d}$ is the node embedding matrix, i.e., row-wise packed pooled feature vectors extracted from salient regions, $C_v \in R^M$ and $C_r \in R^{M \times M}$ are the category matrices of which each element is the index of categories of the corresponding vertices and relationships.

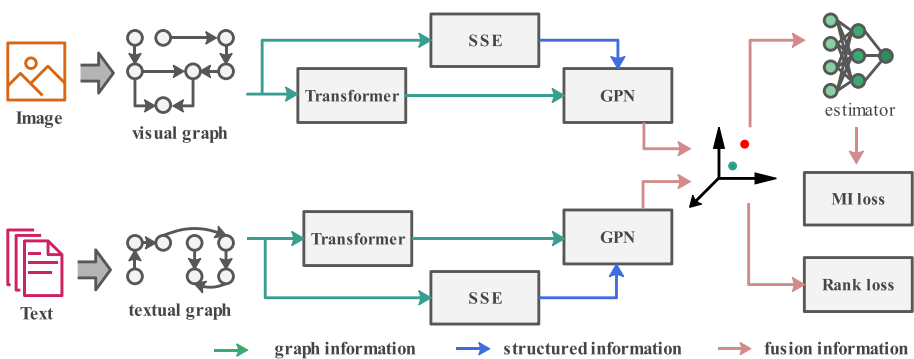


Fig. 1 Overview of the proposed graph attention network

Given the textual data, we first split each sentence to an ordered sequence of words using the WordPiece tokenizer trained in the work [49]. And then we further extract the semantic dependency relationship using the SPICE [50] and Stanford CoreNLP toolkit [51]. Finally, we represent each sentence as a textual semantic graph $\mathcal{G}_T = \{T, C_r\}$, where $T \in R^{N \times d}$ is the sum of word embedding and position embedding. C_r is the category matrix of which each element is an index corresponding to specific dependency relationship. Because the visual relationships are also categorized with respect to various predicates, the C_r in graphs \mathcal{G}_V and \mathcal{G}_T are all the subsets of same relationship set, i.e., the most frequent 50 predicates in Visual Genome dataset. Therefore, it allows that a shared predicate embedding matrix is utilized to mitigate the heterogeneous gap between images and sentences.

3.2 Architecture of network

The proposed graph attention network aims to transform both visual and textual graphs into a common semantic subspace. As illustrated in the Fig. 1, the visual and textual branches are mutually independent and built with similar architecture composed of three components. We will detail the fragment embedding module (FE) in the Section 3.2.1, the structured semantic enhancement (SSE) module in the Section 3.2.2 and graph-based pooling module (GPN) in the Section 3.2.3.

3.2.1 Feature embedding module

As the backbone of our model, the fragment embedding module aims to improve the fragment-level representations for further pooling operation. Due to the prominent performance in deep learning domain, the multi-head attention mechanism [31] is adopted to construct the embedding module. The core of attention mechanism is the scale dot-product attention operation which is defined as the (1)–(2), where X represents the fragment embedding matrices V or T and d_h is the dimension of features in the h -th head. $W_h \in R^{d \times d_h}$ and $b_h \in R^{d_h}$ is the learnable weight matrix and bias.

$$H_h(X) = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_h}}\right) V_h \quad (1)$$

$$Z_h = X W_h^Z + b_h^Z, Z \in \{Q, K, V\} \quad (2)$$

$$\text{Attention}(X) = [H_1(X), \dots, H_h(X)] W_o + b_o \quad (3)$$

Given the outputs of multiple self-attention modules, the multi-head attention mechanism is defined as the (3), where $[\cdot]$ represents the concatenation operation along the feature dimension. $W_o \in R^{d \times d}$ and $b_o \in R^d$ are the learnable weight matrix and bias.

$$\text{FFN}(x) = W_2(\text{GELU}(W_1 x + b_1)) + b_2 \quad (4)$$

$$\tilde{T}(X) = \text{LayerNorm}(X + \text{Attention}(X)) \quad (5)$$

$$X^S = \text{LayerNorm}(\tilde{T}(X) + \text{FFN}(\tilde{T}(X))) \quad (6)$$

Finally, the feature embedding module is define as (4)–(6). The FFN represents a position-wise feed-forward network and the GELU is the Gaussian error linear units [52]. And the normalization layer [53] and residual connection are introduced for improving the stability of training. Note that the feature dimension is invariant due to the residual connection. And we adopt two sequential embedding modules (formulated as (5) and (6)) as the backbones

in visual and textual branches. The first one aims to improve intra-modal semantic features and the second one is combined with the structured semantic enhancement module (detailed in the Section 3.2.2) for generating more informative cross-modal representations.

3.2.2 Structured semantic enhancement module

The structured semantic enhancement (SSE) module aims to embed the relationship label information into the fragment embedding module for further improving the cross-modal semantic alignment. Concretely, we compute a coefficient matrix R given the relationship category matrix $C_r \in R^{M \times M}$ using the (7)-(8). The $W_e^R \in R^{N_R \times d_R}$ is the relationship embedding matrix where the N_R and d_R are respectively the number of relationship categories and dimension of embedding. The \otimes represent a index operation, i.e., replace the elements of C_r with the corresponding row vectors in embedding matrix W_e^R . Then we apply the max pooling operation along the second dimension to aggregate the context label information for each vertex. The weight matrix $W_{ir} \in R^{d \times d_R}$ maps the initial node embeddings into the relationship embedding space. The representations of node features and their context information are concatenated as the relation-aware embeddings. And we further apply a parametric bilinear function (8) to generate the final coefficient matrix R .

$$\tilde{R} = [W_{ir}X, \text{maxpool}[W_e^R \otimes C_r]] \quad (7)$$

$$R_h = \tilde{R}W_{rh}\tilde{R}^T \quad (8)$$

$$H_h(X) = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_h}} + \lambda R_h\right)V_h \quad (9)$$

We embed the structural information into the fragment embedding module by adding the coefficient matrix R to the weight matrix in self-attention mechanism (1) before computing the *softmax* function, which can be rewritten as the (9). The λ is a predefined coefficient. On the one hand, we regard the row vectors of matrix \tilde{R} as the context-aware representations of vertices. The elements of coefficient matrix R is larger if two vertices have more similar context semantic environment. And an important property of *softmax* function is that adding the same scores to each variable dose not change the final weight distribution. Therefore, we leverage the structured information to modulate the fragment embedding module with the summation operation in (9). And it endows the proposed model with the capability that dealing with the graph data while the structured information lacks. Note that we refer the outputs of SSE module as X^R .

3.2.3 Graph-based pooling module

The graph-based pooling module aims to transform the structured semantic enhanced features X^R to a feature point x^s in the shared semantic subspace. To this end, the average and max pooling are widely adopted in the most existing models. However, the average pooling operation treats all feature vectors equally which may suppress the information from which the cross-modal alignment benefits. On the contrary, the max pooling operation preserves the most prominent feature information but may lose some other valuable information. Intuitively, the more other targets an object connects to, the more important it is for understanding the semantic content. Therefore, we argue that introducing structured information in pooling operation is able to preserve as much valuable information as possible. To this end, we

implement the graph-based pooling module with two-layer GCN [33] which is derived from the spectral graph theory.

$$P(X) = X^T \text{softmax}(\tilde{A}(\text{GELU}(\tilde{A}XW_1))W_2) \quad (10)$$

$$\tilde{A} = D^{-1/2}(A)D^{-1/2} \quad \text{where} \quad D_{ii} = \sum_j A_{ij} \quad (11)$$

The GCN treats the graph data as one-dimension signal with multiple channels and filters out the noise and useless information with the parametric filters. The whole pooling function is formulated as the (10) where the $W_1 \in R^{d \times d}$ and $W_2 \in R^{d \times 1}$ are the learnable parameters of graph filters. And the normalized adjacency matrix \tilde{A} is defined as the (11) where the A is the adjacency matrix with self-loop and the D is the diagonal degree matrix. The graph filter takes as input the node features X^S in the (6) and output the weight coefficients used to compute weighted sum of the output of SSE module, i.e., $x^g = P(X)X^R$. Finally, we concatenate the average of self-attention module X^S and the output x^g as the representations of images and sentences.

3.3 Objective functions

In this subsection, we elaborate on the training strategy given the batched N samples $O = \{(v_n, t_n)\}_{n=1}^N$. The $v \in R^d$ and $t \in R^d$ are the visual and textual common representations computed using the (10). We use subscript i and j to index the images and sentences in the batch. We apply the bidirectional triplet rank loss to promote the cross-modal semantic alignment in the shared subspace. The rank loss is defined as the (12) where the s_{ij} is the cosine similarity between image v_i and sentence t_j . The i^- and j^- indicate the hard negative samples [54] in the batch and the $m = 0.2$ is a predefined marginal value. The rank loss aims to push away the irrelevant items and push the relevant items together.

$$L_r = \sum_i [s_{ij^-} - s_{ij^+} + m]_+ + \sum_j [s_{i^-j} - s_{i^+j} + m]_+ \quad (12)$$

We argue that the cross-modal retrieval benefits from the features with high correlation between images and sentences. Therefore, a cross-modal mutual information (MI) estimation loss is used to search the feature subspace with maximum correlation. Follow the work [55], the low bound of mutual information between two high-dimension variables can be estimated using an estimator and the Donsker-Varadhan (DV) representation of KL divergence. Briefly, as the (13), the mutual information between two variables can be formulated as the KL divergence between joint distribution and multiplication of two marginal distributions. Therefore, the DV representation defined as the (14) is the low bound of mutual information, the estimator T is a function from the sample space to the real number space.

$$I(x, y) = P_{x,y} \log \frac{P_{x,y}}{P_x \cdot P_y} = KL(P_{x,y} \| (P_x \cdot P_y)) \quad (13)$$

$$KL(P \| Q) \geq \sup_{T \in \mathcal{F}: \Omega \rightarrow \mathbb{R}} E_P[T] - \log(E_Q[e^T]) \quad (14)$$

In our work, the variables are visual and textual cross-modal representations v and t . We implement the estimator with the neural network defined as (15), where the notation $[\cdot]$ is a concatenation operation. And we compute the final estimated value using a sigmoid function to avoid numerical overflow at the training stage. Therefore, we can define the cross-modal mutual information estimation loss as the (16), where we use the n to index positive pairs and

$(\tilde{v}_n, \tilde{t}_n)$ are the n -th negative sample pairs in the batch. We construct the batch of negative samples by packing each image with the sentence next to its positive sample.

$$M_e(v, t) = W_2(GELU(W_1[v, t] + b_1)) + b_2 \quad (15)$$

$$L_m = \log\left(\frac{1}{N} \sum_n e^{M_e(\tilde{v}_n, \tilde{t}_n)}\right) - \frac{1}{N} \sum_n M_e(v_n, t_n) \quad (16)$$

Finally, the whole objective function is formulated as the $L = \alpha L_r + \beta L_m$, where the α and β are trade-off coefficients. And the proposed graph attention network can be trained end-to-end.

4 Experiments and analysis

4.1 Datasets and evaluation

In this section, we experiment with our model on two publicly available datasets. Flickr30k [56] containing 31,783 images about person and sports. MS-COCO [57] contains 123,287 images which belongs to 91 common object categories. Each image in both Flickr30k and MS-COCO is annotated with five sentences. Follow the work [54, 58], we use same splitting manner for Flickr30k (identical 29000, 1000 and 1000 images for training, validation and testing) and MS-COCO (identical 5000 validation images, 5000 test images and the rest are training images). We evaluate the proposed model using recall rate at K ($R@K$), i.e., the percent of query for which at least one correct sample is returned in the top K retrieved items. Follow the prior work, we report the results on either averaging over five folds of 1000 images or 5000 images for the MS-COCO. We also report the sum of all recall rates in both image and sentence retrieval tasks. The model which achieves the maximum sum of recall rates on validation set is regarded as the optimal model, and the corresponding results on test set are reported.

4.2 Experiment setup

We implement all experiments with the Pytorch [59] framework and optimize the model with AdamW [60]. The dimension d is equal to 768 and the number of attention heads is set to 12. The internal size of feed-forward network (4) is set to 2048. The dimension of category label embedding is set to 512 and the embedding matrix is randomly initialized. The internal size of mutual information estimator is set to 768. For Flickr30k dataset, we train the model with initial learning rate 0.0002 for 30 epochs and decay the learning rate by 10 for every 10 epochs. And the hyper-parameters λ , α and β are set to 0.6, 1.0 and 0.2. For MS-COCO dataset, we train the model with initial learning rate 0.0002 for 40 epochs, and set the learning rate to 0.00002 for the last 20 epochs. And the hyper-parameters λ , α and β are set to 0.6, 1.0 and 0.3. We refer the model with default configuration as FGA in the following experiment results.

4.3 Comparison with state-of-the-art methods

In this section, we compare our model with several state-of-the-art baselines. The fragment-based methods include SCAN [9], SAEM [32], CVSE [42], CAMERA [18] and RRTC [10]. The graph-based methods include VSRN [21], SMFEA [41], ABGR[23] and SGM [24]. The

comparison results on Flickr30k and MS-COCO are respectively illustrated in the Tables 1 and 2. And the superscript * indicate the results are achieved using ensemble technique.

As shown in the Table 1, the proposed model can achieve competitive performance in all tasks. The CAMERA model which achieves the best performance only gains an average 2.5% improvement than our model. However, except for the VSRN, the other models fail to provide the common cross-modal representations for the efficient retrieval. And our model with ensemble technique gains a total 5.0% improvement than the VSRN. On the one hand, we can see that our model achieves more competitive performance in R@10 in both image and sentence retrieval tasks, i.e., only gains an average of 0.65% degradation when compared with the CAMERA. On the other hand, our model gains about 3.4% degradation in both R@1 and R@5. We think the reason lies in the different metric functions adopted by these models. For example, the CAMERA stores each database item as a set of local feature vectors and evaluates the similarity by selecting partial features with respect to the specific query, from which the fine-grained alignment task may benefit. But it also significantly increases the computation burden at the retrieval stage.

Similarly, as shown in the Table 2, our model also achieves competitive performance in all tasks when compared with other models. Due to the balanced category distribution in MS-COCO, the 1000-images retrieval task is easier than in Flickr30k. Therefore, all models achieve similar performance in all 1000-images retrieval tasks. When it comes to the results on 5000-images retrieval task, our model gains only total 0.3% degradation than the VSRN which achieves the best comprehensive performance. And the results on two datasets also shown that the retrieval performance can be further improved using the ensemble technique. In conclusion, our model is capable of achieving acceptable trade-off between efficiency and effectiveness.

4.4 Ablation study

4.4.1 The effect of modules

In this section, we experiment with multiple ablation models to investigate the effectiveness of different components. A baseline model (FGA_{base}) is first constructed with only fragment embedding module and the average pooling operation is adopted to generate the cross-

Table 1 Comparison results with state-of-the-art methods on Flickr30k

Models	Sentence retrieval			Image retrieval			Sum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN	67.4	90.3	95.8	48.6	77.7	85.2	465.0
CVSE	70.5	88.0	92.7	54.7	82.2	88.6	476.7
SAEM	69.1	91.0	95.1	52.4	81.1	88.1	476.8
CAMERA	76.5	95.1	97.2	58.9	84.7	90.2	502.6
RRTC	72.7	93.8	96.8	54.2	79.4	86.1	483.0
VSRN*	71.3	90.6	96.0	54.7	81.8	88.2	482.6
ABGR	72.3	91.8	95.1	53.7	80.1	87.2	480.2
SMFEA	73.7	92.5	96.1	54.7	82.1	88.4	487.5
FGA	71.0	91.7	97.0	54.6	82.1	88.1	484.5
FGA*	71.2	92.3	97.1	55.2	82.8	89.0	487.6

The bold entries refer to the best results in the comparison experiments

Table 2 Comparison results with state-of-the-art methods on MS-COCO

Models	Sentence retrieval			Images retrieval			Sum
	R@1	R@5	R@10	R@1	R@5	R@10	
1000 images							
SCAN	72.7	94.8	98.4	58.8	88.4	94.8	507.9
SGM	73.4	93.8	97.8	57.5	87.3	94.3	504.1
VSRN*	76.2	94.8	98.2	62.8	89.7	95.1	516.8
SAEM	71.2	94.1	97.7	57.8	88.6	94.9	504.3
SMFEA	75.1	95.4	98.3	62.5	90.1	96.2	517.6
CAMERA	75.9	95.5	98.6	62.3	90.1	95.2	517.6
RRTC	76.2	96.3	98.9	61.6	89.3	94.6	516.9
ABGR	73.0	94.7	98.3	59.5	89.4	95.2	510.1
FGA	75.7	95.3	98.7	59.1	89.2	94.2	512.2
FGA*	75.6	95.2	98.8	59.7	90.0	94.4	513.7
5000 images							
SCAN	50.4	82.2	90.0	38.6	69.3	80.4	410.9
SGM	50.0	79.3	87.9	35.3	64.9	76.5	393.9
VSRN*	53.0	81.1	89.4	40.5	70.6	81.1	415.7
CAMERA	53.1	81.3	89.8	39.0	70.5	81.5	415.2
SMFEA	54.2	–	89.9	41.9	–	83.7	–
FGA	52.9	79.4	89.5	39.2	69.7	81.9	412.6
FGA*	52.6	80.9	89.5	40.3	70.0	82.1	415.4

The bold entries refer to the best results in the comparison experiments

modal representations. Then the structured semantic enhancement module is removed from the default model to investigate its effectiveness. We refer the ablation model as FGA_{sse} . Similarly, we construct the model FGA_{gp} by replacing the graph-based module with the average pooling operation. Except for these modules, we also experiment with a model FGA_{mi} which removes the cross-modal mutual information estimation loss function. Finally, we test the model with default configuration using the data which lacks the structured information, i.e., all relationships between fragments. The default value of normalized adjacency matrix \tilde{A} is set to the identity matrix and the coefficient matrix R is set to zero. We refer the model in this case as FGA_{frag} .

We experiment with these models on the Flickr30k dataset and report the results in the Table 3. The default model gains an average 3.9% improvement than the model FGA_{sse} , which proves the effectiveness of structured semantic enhancement module. Similarly, the default model gains an average 2.6% improvement than the model FGA_{gp} , which has shown that the graph-based pooling module works better than an average pooling operation. In contrast to the baseline model, we can see that both enhancement module and pooling module improve the performance to some extent. The baseline model is trained using only fragment-level features. Therefore, it is feasible to promote cross-modal semantic understanding by introducing auxiliary structured information. When it comes to the loss function, the default model gains an average 1.2% improvement than the model FGA_{mi} . The cross-modal mutual information estimation loss slightly improves the retrieval performance by maximizing the low bound of mutual information between visual and textual representations. When the structured information lacks, the model FGA_{frag} performs better than the baseline model by an average 2.3% improvement on all tasks. Because the enhancement module does nothing while the relationship labels lack, the graph-based pooling module mainly contributes to the

Table 3 Results of the ablation experiments on Flickr30k

Models	Sentence retrieval			Image retrieval			Sum
	R@1	R@5	R@10	R@1	R@5	R@10	
FGA	71.0	91.7	97.0	54.6	82.1	88.1	484.5
FGA _{base}	63.9	84.3	90.0	48.2	74.1	78.4	438.9
FGA _{sse}	65.3	86.9	93.4	50.6	77.9	87.0	461.1
FGA _{gp}	67.7	89.2	95.7	52.6	79.6	84.3	469.1
FGA _{mi}	69.1	90.3	96.1	54.0	81.0	86.9	477.4
FGA _{frag}	65.0	85.9	92.1	50.0	77.3	82.6	452.9

improvement. Since we set the default value of adjacency matrix to the identity matrix, the GCN network is equivalent to a two-layer feed forward network given the $\tilde{A} = I$. And the weights of linear maps are the parameters of graph filters. Therefore, the pooling module still tend to preserve as valuable information as possible.

4.4.2 The effect of hyper-parameter

Next we investigate how the performance changes as the value of specific hyper-parameters change. We first experiment with several models of which the trade-off coefficients λ are selected from the range $[0, 1]$ with the interval 0.1. The results on Flickr30k are reported in the Fig. 2(a). We can see that the model achieves approximate performance while the λ ranges from 0.4 to 0.8 and the maximum sum appears at $\lambda = 0.6$. The performance curve shows that introducing the structured information can promote the cross-modal semantic understanding. To further explore the effectiveness of cross-modal mutual information estimation loss, we carry out several experiments in which the values of trade-off coefficient β are selected from the range $[0, 1]$ with the interval 0.1. The results on Flickr30k are reported in the Fig. 2(b). We can see that the best performance appears at $\beta = 0.2$ and the difference between low and upper bounds is approximate 10%, i.e., less than 2.0% on average. In contrast to the other components, the mutual information estimation loss is not capable of significantly affecting the retrieval performance.

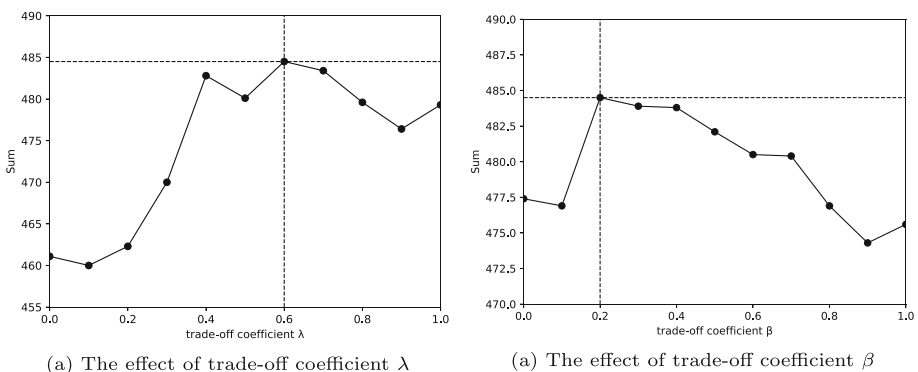
**Fig. 2** The results of ablation experiments on the hyper-parameters

Table 4 Results of the ablation experiments on Flickr30k

Models	Parameters($\times 10^6$)	FLOPs($\times 10^6$)	Dimension	Common space
FGA	32.34	503.01	768	✓
ABGR	110.32	530.67	1024	×
SMFEA	146.30	430.00	1024	✓
SGM	25.28	875.55	1024	×

4.4.3 The effect of graph data

Finally, we investigate the retrieval performance of the model under case where either visual or textual structural data is missing. And we select three classical graph-based cross-modal retrieval models for comparison, i.e., SMFEA [41], ABGR [23] and SGM [24]. We first present a brief comparison between these models in the Table 4 from four aspects: amount of parameters, floating point operations (FLOPs), target dimension and whether the common space exists or not. The first one directly reflects the storage cost of models. And the last three items comprehensively reflect the time cost at both precomputing and retrieval stages. Specially, we set the lengths of image and text sequences to 36 and 1 when computing the FLOPs. We can see that our model has considerable advantage in time and storage cost.

We carry out the comparison experiments on Flickr30k dataset and train our model with the default configuration mentioned above. The other models are trained with the configuration reported in the original papers and open-source codes. All models adapt to the graph data preprocessed as described in the Section 3.1. We report all results in the Table 5. We refer the complete input data as visual graph G_I and textual graph G_C , the incomplete data as V_I and V_C . For our model, we set the default adjacent matrix R to the identity matrix when the structural information is missing. For other models, we simply treat the incomplete data as the graph with only self-loop and represent the self-loop with background label embedding. Consider that the SMFEA model takes as input the graph with fixed three-layer structure, we appropriately trim our scene graph to meet the requirement. From the results of both scenes, we can see that our model performs better than the other models. Obviously, the incomplete data is not taken into account when the other models are designed. And the model achieves better comprehensive performance in the second scene in the Table 5. In our view, it may be caused by the fact that exact scene graph is easier to detect for text than image.

Table 5 Results of the ablation experiments on Flickr30k

Scene	Models	Sentence retrieval			Image retrieval			Sum
		R@1	R@5	R@10	R@1	R@5	R@10	
G_I, V_C	FGA	64.7	86.0	92.3	51.0	77.6	83.1	454.7
	ABGR	60.1	78.7	86.0	42.3	69.3	74.5	410.9
	SMFEA	62.3	81.1	88.3	43.1	70.6	78.1	423.5
	SGM	61.2	80.4	86.7	43.0	72.1	78.3	421.7
V_I, G_C	FGA	67.2	88.5	94.0	51.1	77.0	83.4	461.2
	ABGR	59.3	76.7	82.1	38.3	65.8	70.3	392.5
	SMFEA	63.0	80.9	88.1	40.2	67.5	77.6	417.3
	SGM	63.7	83.5	89.9	45.5	75.0	80.1	437.7

4.5 Visualization analysis

To better illustrate the effect of the proposed method, we visualize the learned cross-modal representations in a 2-D subspace using the t-SNE algorithm [61] in the Fig. 3. We randomly select 100 images and 500 corresponding sentences from test set in Flicikr30k for visualization analysis, and visualize the distribution of these learned common representations in the Fig. 3. Image and sentence are respectively marked as product sign and dot, the points with same color represent those samples belonging to the same semantic category. From the visualized result in Fig. 3, we can see that the most of samples are located near the semantically relevant items. And there is no obvious trend that similar samples are distributed along the radial direction.

Next we provide several retrieval examples selected from the test set in MS-COCO to illustrate the effect of the proposed model. The results of sentence retrieval are illustrated in the Table 6, where the top-5 sentences are arranged from top to bottom in order. And we report the negative sentences using red font. We can see that all positives samples are ranked at top 5 for the first two examples, and the negative samples in the rest examples also present partial relevant semantic information. Table 7 shows several examples of image retrieval and the samples are ranked from left to right. We mark the positives samples using green bound box. We can see the correct image can be ranked at top 1 for the first three examples and those incorrect samples generally present similar visual semantic information.

5 Conclusion

Many prominent works in multimodal deep learning domain have shown that structured graph information is capable of improving semantic understanding. In this paper, we attempt to learn the common cross-modal representations for heterogeneous graph data, i.e., project both visual and textual graphs into a shared semantic subspace. To this end, we propose a graph attention network to embed structured semantic information into the learned cross-modal representations. Concretely, a structured semantic enhancement module leverages the structured information to modify the attention weights for fragment-level feature enhance-

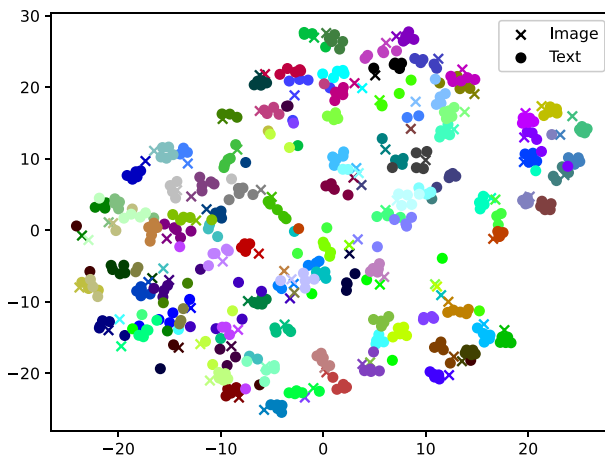


Fig. 3 Visualization of the cross-modal representations using t-SNE algorithm

Table 6 The examples of sentence retrieval


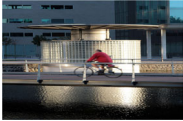


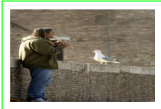











Image query	Results
	Dog with orange ball at feet, stands on shore shaking off water. A white dog shakes on the edge of a beach with an orange ball. White dog playing with a red ball on the shore near the water. A dog shakes its head near the shore , a red ball next to it. White dog with brown ears standing near water with head turned to one side.
	A man in a red long-sleeved shirt bikes over a body of water on a bridge. A man riding his bike across the bridge that is over the river. A man in red on a bicycle rides past a glass structure. Man in a red shirt riding his bicycle around water. A man in a red shirt rides his bicycle.
	Two different breeds of brown and white dogs play on the beach. Two dogs run towards each other on a rocky area with water in the background. Two large tan dogs play along a sandy beach. Two dogs playing together on a beach. A white dog is running down a rocky beach.
	Yellow-orange pecial purpose train engine with American flag painted in the side. A very bright colored train near a big building. A group of yellow and blue umbrellas near a building clock. A lot of blue and yellow umbrellas sitting under a clock. Many blue and yellow umbrellas are shown next to a building.

Table 7 The examples of image retrieval

Text query	Results
A long-haired man is playing the recorder and a seagull is sitting nearby on a wall.	  
A girl with a soccer ball at her feet is standing in front of a boy.	  
A bathroom featuring a walk in shower, mirror, sink and toilet.	  
A lone climber on a snowcapped mountain with several huge mountains in the background.	  

ment. And a graph-based pooling module compresses a set of enhanced features to a single vector in the shared semantic subspace.

In contrast to the most of existing fragment-level and graph-level methods, the proposed model is capable of constructing an explicit semantic subspace where retrieval is equivalent to the vector-based ranking problem, which make it feasible to process a huge amount of data with acceptable time cost. And the proposed model is flexible to process the data no matter whether there exists structured semantic information. The comparison experiments on two publicly available datasets show that our model achieves competitive performance when compared with several state-of-the-art models. And the ablation experiments have also shown that our model is capable of achieving effective retrieval while the structured information lacks.

However, there are still some shortcomings in our work. One is that the work of graph-based pooling module is to generate the weight distribution, i.e., the common cross-modal representations are still essentially the linear combination of fragment-level features, which may limit the quality of cross-modal representations. The other one is that our model still need to select some empirical hyper-parameters for various datasets. To mitigate these problem, the future work will attempt to study a more robust and superior strategy to embed structured semantic information into the cross-modal representations.

Acknowledgements The research was supported by The National Natural Science Foundation of China (grant nos. 61728204, 61802182).

Data Availability The Flickr30k and MSCOCO are publicly available datasets in published papers [56, 57]. The code of the current study are available from the corresponding author on reasonable request.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Liu D, Cui Y, Cao Z, Chen Y (2020) Indoor navigation for mobile agents: a multimodal vision fusion model, pp 1–8
2. Yan L, Liu D, Song Y, Yu C (2020) Multimodal aggregation approach for memory vision-voice indoor navigation with meta-learning 5847–5854
3. Yan L et al (2022) Gl-rg: global-local representation granularity for video captioning
4. Wang Q et al (2022) Webformer: the web-page transformer for structure information extraction. Proc ACM Web Conf 2022:1–2
5. Yang L et al (2023) Findings of the association for computational linguistics. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Mixpave: mix-prompt tuning for few-shot product attribute value extraction: ACL
6. Wang Q et al (2023) Proceedings of the 61st annual meeting of the association for computational linguistics. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Mustie: multimodal structural transformer for web information extraction, vol 1. Association for Computational Linguistics, Toronto, pp 2405–2420
7. Huang Y, Wang W, Wang L (2017) Instance-aware image and sentence matching with selective multimodal lstm 1:2310–2318
8. Nam H, Ha J-W, Kim J (2017) Dual attention networks for multimodal reasoning and matching, pp 299–307
9. Lee K-H, Chen X, Hua G, Hu H, He X (2018) Stacked cross attention for image-text matching, pp 201–216
10. Wu J, Wu C, Lu J, Wang L, Cui X (2022) Region reinforcement network with topic constraint for image-text matching. IEEE Trans Circuits Syst Video Technol 32(1):388–397. <https://doi.org/10.1109/TCSVT.2021.3060713>
11. Zhang K, Mao Z, Wang Q, Zhang Y (2022) Negative-aware attention framework for image-text matching. IEEE, pp 15640–15649. <https://doi.org/10.1109/CVPR52688.2022.01521>

12. Wang S, Chen Y, Zhuo J, Huang Q, Tian Q (2018) Joint global and co-attentive representation learning for image-sentence retrieval, pp 1398–1406
13. Zhang Q, Lei Z, Zhang Z, Li SZ (2020) Context-aware attention network for image-text retrieval, pp 3536–3545
14. Yu T et al (2021) Heterogeneous attention network for effective and efficient cross-modal retrieval, pp 1146–1156
15. Wei X, Zhang T, Li Y, Zhang Y, Wu F (2020) Multi-modality cross attention network for image and sentence matching, pp 10941–10950
16. Qu L, Liu M, Wu J, Gao Z, Nie L (2021) Dynamic modality interaction modeling for image-text retrieval, pp 1104–1113
17. Li J, Niu L, Zhang L (2022) Action-aware embedding enhancement for image-text retrieval. AAAI Press 1:1323–1331
18. Qu L, Liu M, Cao D, Nie L, Tian Q (2020) Context-aware multi-view summarization network for image-text matching, pp 1047–1055
19. Lan H, Zhang P (2022) Learning and integrating multi-level matching features for image-text retrieval. IEEE Signal Process Lett 29:374–378. <https://doi.org/10.1109/LSP.2021.3135825>
20. Cheng Z et al (2023) Fusion is not enough: single-modal attacks to compromise fusion models in autonomous driving. [ArXiv:abs/2304.14614](https://arxiv.org/abs/2304.14614), <https://api.semanticscholar.org/CorpusID:258417952>
21. Li K, Zhang Y, Li K, Li Y, Fu Y (2019) Visual semantic reasoning for image-text matching, pp 4654–4662
22. Liu C et al (2020) Graph structured network for image-text matching, pp 10921–10930
23. Zhong, X et al (2021) Auxiliary bi-level graph representation for cross-modal image-text retrieval. IEEE, pp 1–6
24. Wang S, Wang R, Yao Z, Shan S, Chen X (2020) Cross-modal scene graph matching for relationship-aware image-text retrieval, pp 1508–1517
25. Long S, Han SC, Wan X, Poon J (2022) Gradual: graph-based dual-modal representation for image-text matching. IEEE, pp 2463–2472. <https://doi.org/10.1109/WACV51458.2022.00252>
26. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
27. Huang F, Zhang X, Zhao Z, Li Z (2019) Bi-directional spatial-semantic attention networks for image-text matching. IEEE Trans Image Process 28(4):2008–2020. <https://doi.org/10.1109/TIP.2018.2882225>
28. Ji Z, Wang H, Han J, Pang Y (2019) Saliency-guided attention network for image-sentence matching, pp 5754–5763
29. Liu C et al (2019) Focus your attention: a bidirectional focal attention network for image-text matching, MM '19. Association for Computing Machinery, New York, pp 3–11
30. Wang Y et al (2019) Position focused attention network for image-text matching. [arXiv:1907.09748](https://arxiv.org/abs/1907.09748)
31. Vaswani A et al (2017) Attention is all you need 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
32. Wu Y, Wang S, Song G, Huang Q (2019) Learning fragment self-attention embeddings for image-text matching, pp 2088–2096
33. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks, OpenReview.net.<https://openreview.net/forum?id=SJU4ayYgl>
34. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder-decoder approaches. [arXiv:1409.1259](https://arxiv.org/abs/1409.1259)
35. Zellers R, Yatskar M, Thomson S, Choi Y (2018) Neural motifs: scene graph parsing with global context, pp 5831–5840
36. Li Y, Ouyang W, Zhou B, Wang K, Wang, X (2017) Scene graph generation from objects, phrases and region captions, pp 1261–1270
37. Krishna R et al (2016) Visual genome: connecting language and vision using crowdsourced dense image annotations. [arXiv:1602.07332](https://arxiv.org/abs/1602.07332)
38. Xu D, Zhu Y, Choy CB, Fei-Fei L (2017) Scene graph generation by iterative message passing, pp 3097–3106
39. Liang Y et al (2019) Vrr-vg: refocusing visually-relevant relationships. IEEE, pp 10402–10411. <https://doi.org/10.1109/ICCV.2019.01050>
40. Lu X, Zhu L, Liu L, Nie L, Zhang H (2021) Graph convolutional multi-modal hashing for flexible multimedia retrieval, pp 1414–1422. <https://doi.org/10.1145/3474085.3475598>
41. Ge X et al (2021) Structured multi-modal feature embedding and alignment for image-sentence retrieval. ACM, pp 5185–5193
42. Wang H, Zhang Y, Ji Z, Pang Y, Ma L (2020) Consensus-aware visual-semantic embedding for image-text matching. Lecture notes in computer science, vol 12369. Springer, pp 18–34
43. Yan L, Cui Y, Chen Y, Liu D (2021) Hierarchical attention fusion for geo-localization, pp 2220–2224

44. Cui Y, Yan L, Cao Z, Liu D (2021) Tf-blender: temporal feature blender for video object detection, pp 8138–8147
45. Liu D et al (2021) Densernet: weakly supervised visual localization using multi-scale feature aggregation. *Proc AAAI Conf Artif Intell* 35(7):6101–6109. <https://doi.org/10.1609/aaai.v35i7.16760>
46. Cui Y et al (2021) Geometric attentional dynamic graph convolutional neural networks for point cloud analysis. *Neurocomputing* 432:300–310
47. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. *Adv Neural Inf Process* 28:91–99
48. Tang K, Niu Y, Huang J, Shi J, Zhang H (2020) Unbiased scene graph generation from biased training. *Computer Vision Foundation/IEEE* 1:3713–3722
49. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding
50. Anderson P, Fernando B, Johnson M, Gould S (2016) Computer vision – ECCV 2016. In: Leibe B, Matas J, Sebe N, Welling M (eds) *Spice: semantic propositional image caption evaluation*. Springer International Publishing, Cham, pp 382–398
51. Manning CD et al (2014) The stanford corenlp natural language processing toolkit. *Assoc Comput Linguistics* 1:55–60. <https://doi.org/10.3115/v1/p14-5010>
52. Hendrycks D, Gimpel K (2016) Gaussian error linear units (gelus)
53. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. [arXiv:1607.06450](https://arxiv.org/abs/1607.06450)
54. Faghri F, Fleet DJ, Kiros JR, Fidler S (2017) Vse++: improving visual-semantic embeddings with hard negatives. [arXiv:1707.05612](https://arxiv.org/abs/1707.05612)
55. Belghazi MI et al (2018) Mutual information neural estimation. (eds Dy JG, Krause A) *Proceedings of the 35th international conference on machine learning, ICML 2018, vol 80. Proceedings of Machine Learning Research, Stockholm*, pp 530–539. <http://proceedings.mlr.press/v80/belghazi18a.html>
56. Plummer BA et al (2015) Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models, pp 2641–2649
57. Lin T-Y et al (2014) Microsoft coco: common objects in context. Springer, 740–755
58. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions, pp 3128–3137
59. Collobert R, Kavukcuoglu K, Farabet C (2011) Torch7: a matlab-like environment for machine learning
60. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. [arXiv:1711.05101](https://arxiv.org/abs/1711.05101)
61. van der Maaten L, Hinton GE (2008) Visualizing data using t-sne. *J Mach Learn Res* 9:2579–2605

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.