# Learning Dual Semantic Relations with Graph Attention for Image-Text Matching

Keyu Wen, Xiaodong Gu, and Qingrong Cheng

Abstract-Image-Text Matching is one major task in crossmodal information processing. The main challenge is to learn the unified visual and textual representations. Previous methods that perform well on this task primarily focus on not only the alignment between region features in images and the corresponding words in sentences, but also the alignment between relations of regions and relational words. However, the lack of joint learning of regional features and global features will cause the regional features to lose contact with the global context, leading to the mismatch with those non-object words which have global meanings in some sentences. In this work, in order to alleviate this issue, it is necessary to enhance the relations between regions and the relations between regional and global concepts to obtain a more accurate visual representation so as to be better correlated to the corresponding text. Thus, a novel multi-level semantic relations enhancement approach named Dual Semantic Relations Attention Network(DSRAN) is proposed which mainly consists of two modules, separate semantic relations module and the joint semantic relations module. DSRAN performs graph attention in both modules respectively for region-level relations enhancement and regional-global relations enhancement at the same time. With these two modules, different hierarchies of semantic relations are learned simultaneously, thus promoting the image-text matching process by providing more information for the final visual representation. Quantitative experimental results have been performed on MS-COCO and Flickr30K and our method outperforms previous approaches by a large margin due to the effectiveness of the dual semantic relations learning scheme. Codes are available at https://github.com/kywen1119/DSRAN.

 ${\it Index~Terms} \hbox{--} cross\hbox{--} modal~retrieval,~graph~attention,~semantic~relation,~image~text~matching}$ 

#### I. INTRODUCTION

ITH the rapid development of information technology, people's daily life is full of data of various modalities, so cross-modal information processing is increasingly important. For all information, visual and textual forms occupy a dominant position, thus attracting researchers to focus on cross-modal practical tasks of vision and language. For example, the most compelling three hotspots are cross-modal retrieval [1]–[4], visual question answering [5], [6] and image captioning [7], [8]. In this paper, we concentrate on a subtask of cross-modal retrieval that focuses on images and texts named image-text matching. Image text matching can be studied as an independent task or as a solution to other upper level tasks. For example, Huang et al. [9] adopted image-text

This work was supported in part by National Natural Science Foundation of China under grants 61771145 and 61371148. (Corresponding author: Xiaodong Gu. Email: xdgu@fudan.edu.cn)

The authors are with Department of Electronic Engineering, Fudan University, Shanghai 200433, China. Email: kywen19@fudan.edu.cn.

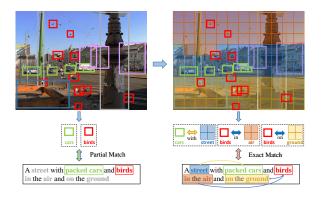


Fig. 1. The proposed DSRAN learns semantic relations between regional objects as well as objects and global context. Salient objects are marked with different colored boxes. Grid shaped shadow in the right picture denotes global context such as "air", "street" and "ground". With only regional features, the visual representations fail to match the corresponding words and relations like "birds in the air". "birds on the ground" or "street with cars".

matching to do visual-textual reasoning for recognizing crossmedia entailment (RCE). By matching specific regions and words, regions in the raw image are assigned with different labels, which benefits fine-grained visual categorization [10]. Different from traditional single-modal retrieval, image-text matching [11] requires the retrieval from image to text and vice versa, which is to find the most relevant text given the query image named image-based text retrieval or to find the semantically most similar image with the query text which is text-based image retrieval.

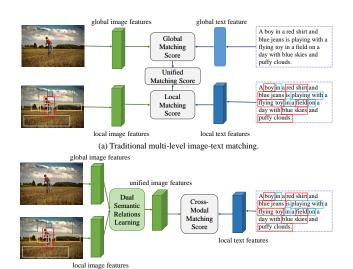
The core challenge for image-text matching is to learn a subspace where the similarities of encoded images and text representations can be directly computed, and the similarities of most correlating image-text pairs are maximized. To begin with, canonical correlation analysis (CCA) [1] stands as a backbone for this task by using a linear projection to project images and texts into the subspace. Recently with the development of deep neural networks, [2], [3], [12], [13] utilize DNN to do this task. Traditionally image and text inputs are separately encoded into global visual or textual features by convolution-based networks [14]–[16] and RNNbased networks like LSTM [17] or GRU [18], after which a similarity function is used to measure the distance between two-modal representations. More recently, text representations can be obtained with pre-trained transformer-based models like BERT [19]. Pre-trained language models contain prior semantic information, which is comparable to the pre-trained CNNs in the image channel. At last, a triplet-based ranking loss function [3] supervises the training and the best unified latent space is learned. Since these methods only take global

features into account, concerning only the alignment of the whole image and the sentence, they lack a more detailed match, that is, a match between the image patches and the words in sentences.

A more refined way is to extract the local regional features using object detection methods such as Faster R-CNN [20], which is called bottom-up attention [6] for cross-modal tasks. With a pre-trained Faster-RCNN, objects in an original image can be detected. Regional features are extracted from these objects by the backbone CNNs like ResNet101 [16]. SCAN [21] firstly introduces this scheme into the image-text matching task and designs stacked cross-modal attention to align the regions in images and words in sentences. The image-text similarities are integrated from all the region-word pairs. VSRN [22] takes a step further to learn the relations between objects in the raw image using graph convolutional network. Since relations learning shows an important status in image-text matching, scene-graph based methods [23]-[25] introduce scene graph generator to generate visual and textual scene graphs for better phrase matching. Although quite successful, these methods seem deficient at two important points: a) The alignment of regions and words seems too single for image-text matching, while the global matching process should be considered. b) They lack emphasis on relations between objects and nonobject elements like the background, the surroundings, or the environment which have a strong relation to the understanding of an image when trying to match the corresponding text. As illustrated in Fig. 1, without learning such relations, only salient objects "birds" and "cars" are matched with the caption. However, learning the relations between objects and global context helps to assign more detailed attributes to the salient objects. For example, by adding the global-region relation "in the air" to the original object "birds", the visual representation can exactly match its textual counterpart.

For problem a, multi-level matching methods alleviate this by integrating both local similarities and global similarities [26]–[30], as shown in Fig. 2 (a). They either boost the combination of image and text information by extracting both local and global features of images and captions and learning their similarities simultaneously or use a multi-label extraction scheme to get global concepts for global alignment based on regional matching methods. However for problem b, whether for global, regional, or multi-level matching methods, they all miss the attention on relations between regional objects and global concepts.

Thus, based on previous work, this paper proposes a Dual Semantic Relations Attention Network (DSRAN) to address this problem. Intuitively there are two main modules in DSRAN, the separate semantic relations module and the joint semantic relations module. The separate semantic relations module is designed for capturing regions' semantic relations. Specifically, because of the efficiency and effectiveness of GATs [31] when learning the nodes relations, the module uses two separate graph attention networks to learn pixel-wise semantic relations and regional relations at the same time. The second module, the joint semantic relations module, aims to find the semantic relations between local objects and global pixel-wise concepts. A unified graph attention network is used



(b) Our proposed multi-level image-text matching based on dual semantic relations learning.

Fig. 2. Different architectures of (a) traditional multi-level image-text matching methods and (b) our proposed multi-level image-text matching based on dual semantic relations learning. Objects and relations in the caption are highlighted with red and blue boxes.

to achieve this. After these two principal relation-oriented modules, a gated fusion process helps to select more useful information for the final visual representation. In the end, the similarity scores of the obtained image features and text features can be calculated for further computing the value of loss function and updating the network parameters as previous works did. Details will be discussed later.

To verify our proposed model's validity, we test our model on both MSCOCO [32] and Flickr30K [33] datasets. Experimental results show that our model outperforms the current state-of-the-art methods on both datasets, proving the effectiveness of our design.

Our contributions are summarized below.

- (a) We propose a novel Dual Semantic Relations Attention Network(DSRAN) in order to strengthen the relations between regional objects and global concepts in the learned visual representations while considering the relations among objects themselves at the same time.
- (b) We propose a new way to learn both global and local consistency by learning a global-regional unified visual representation, instead of learning global and local similarities respectively.
- (c) The proposed DSRAN outperforms previous works on the image-text matching task. Specifically, on MSCOCO our model outnumbers the current best model VSRN [22] by 3.0% for image retrieval and 9.2% for text retrieval (Recall@1 using 5K test set). Moreover, on Flickr30K, the increase is more significant, which is 8.2% for image retrieval and 12.9% for text retrieval (Recall@1).

# II. RELATED WORK

# A. Global Image-Text Matching

Image text matching can mainly be divided into three kinds: global matching methods, regional matching methods and multi-level matching methods. Specifically, for global imagetext matching methods, the goal is to embed raw images and

texts into a common subspace in an end-to-end way, where similarities of the embedded visual and textual features can be directly calculated. The primary challenge lies in the respective mapping progress of images and texts.

Initially, CCA [1] utilizes a linear projection to encode cross-modal data into a common subspace where they are highly correlated. Later, researchers apply DNNs into the projection process like [12], [13], [34]. DCCA [12] constructs multiple stacked layers of nonlinear transformation and learns the maximized correlations of visual and textual representations. Further in DSPE [34], correlation learning between cross-modal encoded features is enhanced by constructing a triplet ranking loss. However, these methods seem too plain to extract the abundant information in the images and captions. Thus, Kiros et al. [13] introduced the CNN-LSTM structure to learn a joint image-text embedding. Because of the strong ability convolutional neural networks show in image processing, a CNN pre-trained on the ImageNet dataset [35] takes the original image as input and outputs the encoded image features. Samely, LSTM [17] and GRU [18] show powerful strength in natural language processing thus they are used to extract the global semantic features of the sentences. VSE++ [3] introduces the concept of hard negatives, which act as a basis for many subsequent studies. With the success of pretraining in NLP field like BERT [19] and GPT [36], more specific text representations can be learned as did in TOD-Net [37].

Inspired by generative adversarial networks [38], similar generative and adversarial learning schemes can be applied in image text matching task by reducing the heterogeneous gap between two modalities. CM-GANs [39] and ACMR [2] add a modality discriminator to the traditional two-way network to distinguish the modal information of the features. When unable to judge, it is considered that the heterogeneous gap between the two modalities has been eliminated. GXN [40] generates images or captions using the learned textual or visual features, thus boosting reducing cross-modal information gap. Wen et al. [41] proposed a cross memory network with pair discrimination to capture the common knowledge between image and text modalities.

More special mechanisms are used in the global-wise matching. DAN [42] applies attention mechanisms for both visual and textual features enhancement. In MTFN [43], Wang et al. proposed a re-ranking scheme for a more precise ranking process during testing. To achieve more comprehensive matching, MFM [44] utilizes multi-faceted representations of image and text to characterize them more comprehensively. Thus the matching relationship between two modalities is discovered from multiple perspectives. Ji et al. [45] proposed Saliency-guided Attention Network (SAN) which adopts visual saliency detection to highlight visually salient regions or objects in an image in accordance with words in sentences.

All the methods mentioned above are summarized as the global image-text matching methods as they all directly encode the whole images or texts into vectors. However, for these methods, they simply consider the alignment of the global context of images or texts while ignoring the alignment of image regions and words, which can be alleviated by the

regional image-text matching methods as described below.

# B. Regional Image-Text Matching

A more refined way for matching images and texts is to match the salient regions in the images and the words in the sentences rather than merely matching the global semantics. Instead of traditional CNNs for image feature extraction, regional image-text matching methods utilize object detection [20] to detect the objects in the images. At the same time, the text encoder no longer outputs the global sentence vectors but word-level matrices.

Thus methods like [11], [21], [22], [46] take the alignment of image regions and words as the alignment of the whole image and sentence. Karpathy et al. [11] first came up with a way to detect objects in images and encoded them into the subspace, where pair-wise image-text similarity is calculated by summing up similarities of all region-word pairs. SCAN [21] introduces bottom-up attention [6] scheme and uses a Faster R-CNN [20] pre-trained on Visual-Genome dataset [47] to encode images into region-level features while texts are encoded into word-level features. Then stacked cross-modal attention is used for similarity calculation. PFAN [46] goes a step further by adding position information of regions into account and designs a position information integration scheme for better matching. Considering there are always words referring to relations in the sentences, VSRN [22] applies GCNs [48] to do the visual reasoning and learns the relations among regions that are in accord with the text modality. Hu et al. [49] proposed a relation-wise dual attention network (RDAN) to capture multi-level cross-modal relations with a visual-semantic relation CNN model and apply dual pathway cross-modal attention. Knowing the importance of learning relations between regions and between words for more refined matching, researchers construct scene graphs for images and sentences separately in which regions/words and relations are encoded into an exclusive graph as in [23]-[25]. Thus the matching process of images and texts turns into the matching of visual scene graphs and textual scene graphs which is more subtle. In [50], the authors extended the supervised image-text matching to an unsupervised one using domain adaptation with scene graph.

#### C. Multi-Level Image-Text Matching

Either the global matching methods or the regional matching methods seem too single for the fine-grained matching which requires not only the alignment of global context but also the alignment of regional objects and words as well as the alignment of relations. The core concept of multi-level matching methods is to learn the correlation of both the global context and the regional/word-level concept.

An easily thought of way is to deal with the problem in a multi-pathway as in [26]–[29]. In these works, global similarity and local similarity are calculated in separate paths and are integrated into the final similarity, as shown in Fig. 2 (a). Specifically in MDM [28] and GSLS [27], two separate paths are designed for respective global and local similarity calculations. They use either a local image CNN or a Faster

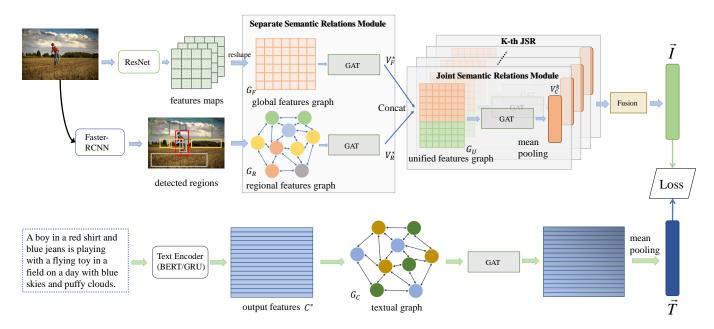


Fig. 3. An overview of the proposed Dual Semantic Relations Attention Network. The model captures dual-path image features with ResNet and Faster R-CNN. Nodes of  $G_F$  and  $G_R$  are global regional features and local object features respectively. Two semantic relations learning modules are applied to enhance both object-level relations and unified global-regional relations. A separate or unified graph attention network is used for each of the two modules. The caption is fed to whether a pre-trained BERT-BASE encoder or a GRU encoder and we construct the textual graph with the output features where nodes of the graph are word representations. Furthermore, a textual GAT is designed for deeper textual relations learning.

RCNN to encode the local features, and use a CNN to encode global features. For texts, the original sentences are encoded into either global feature vectors or word-level matrices. Further, in BSSAN [29] the word to regions (W2R) attention network and the object to words (O2W) attention network are added to each path to compute attention-based matching scores. Additionally, CRAN [26] expands the network to a three-paths one by adding the relation matching path, where the relations of image patches and of words are extracted and matched.

Other methods define the global context in a more specific way. For example, Xu et al. [30] dealt with the global alignment of images and texts in a different way. In their model, local similarity calculation is the same as that in SCAN [21]. They extract semantic labels of images and sentences as the global context with a multi-label classification module for global consistency.

Although these works successfully learn multi-level consistency of images and texts for their matching, they have two apparent deficiencies: a) The design of computing global and local similarity separately limits the model to learn the relations between local objects and some global information, as illustrated in Fig. 1. b) Most of these works compute the final similarity by integrating global similarity, local similarity and even relation similarity, which seems too complicated for real-time applications, for it increases not only calculation complexity but also memory usage. Our work differs from previous multi-level matching methods just in the two points mentioned before. As illustrated in Fig. 2, our DSRAN encodes the global and local features at the same time, learns the relations among themselves or between them, and obtains the visual representation with enhanced relations. Then the model

calculates the similarity of the two-modal features directly and then matches them.

# D. Graph Attention Network

In this paper, we apply graph attention networks [31] to capture both object-level visual relations and global-regional visual relations. As much visual information can not always be expressed as a grid-like structure such as the visual graphs, GNNs [51] is first introduced as a generalization of recursive neural networks which can directly process the graphs. Then GCN [48] was proposed and further utilized in visual relations capturing as did in VSRN [22]. Recently an attention-based method directly using at graphs named GAT [31] overcomes disadvantages of GCN with masked self-attention mechanisms. By stacking the layers that nodes can participate in their neighborhood features, different weights can be implicitly assigned to different nodes in the neighborhood, without any type of computation-intensive matrix operation or prior understanding of the graph structure. RE-GAT [52] introduces this kind of attention based graph networks to do visual relational reasoning and promotes the cross-modal information learning in VQA.

# III. PROPOSED METHOD

## A. Overview of Our DSRAN Approach

In this section, we detail our proposed Dual Semantic Relations Attention Network(DSRAN). As shown in Fig. 3, given an image-text pair, two separate encoding paths are designed for two modalities to get the final representations. For the image part, the raw image is firstly extracted in two levels, the global level and the object level (III-B). Two modules are

followed, the first of which is the separate semantic relations module aiming to learn the region-level semantic relations (III-C). The second is the joint semantic relations module, which is designed for capturing relations across objects and global concepts (III-D). For the text part, either a pre-trained BERT-BASE model [19] or a GRU encoder [18] extracts the representations of the words corresponding to the image features (III-E). With the cross-modal representations, we can calculate the similarity scores and update the network parameters with the loss function(III-F). In the end, we use a re-ranking process for more refined matching(III-G).

# B. Two Levels of Image Features

Given a raw image I, global-level features F and region-level features R are extracted respectively. Generally, a ResNet152 [16] pretrained on ImageNet [35] whose last fully-connect layer is removed extracts the global features of the image. We use the feature map of last layer and reshape it to a set of features  $F = \{f_i, ..., f_n\}, f_i \in \mathbb{R}^{D_o}$  where n is the reshaped feature map size and  $D_o$  refers to the dimension of each pixel. For the region-level part, inspired by bottom-up attention [6], the objects are firstly detected by a Faster-RCNN [20] pretrained on Visual-Genome [47] dataset and then fed into a backbone Resnet101 and the output features can be represented as  $R = \{r_i, ..., r_k\}, r_i \in \mathbb{R}^{D_o}$  where k is the detected objects number. In order to embed them into the shared latent space, a fully-connect layer is carried out.

$$V_F = W_f F + b_f, V_R = W_r R + b_r.$$
 (1)

 $W_f$  and  $W_r$  are the weight matrices together with the bias  $b_f$  and  $b_r$ . Then we get the two-levels extracted features  $V_F \in \mathbb{R}^{D_e}$  and  $V_R \in \mathbb{R}^{D_e}$  representing visual global features and regional features where  $D_e$  is the embedding dimension.

#### C. Separate Semantic Relations Module

Targeting at dual-level features, we design two separate semantic relations enhancement models for learning the enhanced pixel-wise relations and object-wise relations. Specifically, we detail them in three parts, the first of which is the construction of the graph attention module.

# • Graph Attention Module

Given a fully-connected graph G = (V, E), where  $V = \{v_i, ..., v_N\}, v_i \in \mathbb{R}^D$  is the node features and E is the edge set. Following [31], we compute attention coefficients and normalize them with softmax function.

$$e_{ij} = a(W_a v_i, W_k v_i), \tag{2}$$

$$a(W_q v_i, W_k v_j) = W_q v_i (W_k v_j)^T / \sqrt{D}, \tag{3}$$

$$\alpha_{ij} = Softmax(e_{ij}). \tag{4}$$

 $W_q$  and  $W_k$  are learnable parameters. In case of memory explosion, different from using the feed-forward neural network as did in the original GAT [31], we compute the attention

coefficients with multi-head dot production [53], which is much faster and more space-efficient in practice.

$$MultiHead(v_i, v_j) = W_o|_{h=1}^{H}(head_1, ..., head_h),$$
 (5)

where

$$head_h = Softmax(\frac{W_q^h v_i (W_k^h v_j)^T}{\sqrt{d}}) W_v^h v_j.$$
 (6)

In eq.(5),  $\parallel$  means concatenation. The projections are parameter matrices  $W_q^h \in \mathbb{R}^{D \times d}$ ,  $W_k^h \in \mathbb{R}^{D \times d}$ ,  $W_v^h \in \mathbb{R}^{D \times d}$  and  $W_o \in \mathbb{R}^{D \times D}$ .

In this paper, we employ H = 8 parallel attention layers thus d equals D/8. Then, with a nonlinear activation function, the final output feature can be computed.

$$v_{i}^{'} = ReLU(\sum_{j \in N_{i}} MultiHead(v_{i}, v_{j})). \tag{7}$$

Here  $N_i$  is the neighborhood of node i in the graph. We add a batch normalization into the graph attention module to accelerate training.

$$v_i^{'} = BN(v_i^{'}). \tag{8}$$

In eq.(8) BN is the batch normalization layer. Here we finish the construction of a graph attention module.

# • Attention for pixel-wise relations enhancement

Obtaining the global features  $V_F$ , firstly we construct the global visual graph  $G_F = (V_F, E_F)$ . Here the edge set  $E_F$  is defined as the affinity matrix by calculating the affinity edge of each pair of global features  $v_F^i$  and  $v_F^j$ ,

$$E_F(v_F^i, v_F^j) = (v_F^i)^T v_F^j. (9)$$

Here more correlated image regions have edges of higher affinity scores. Thus we have the fully-connected global visual graph  $G_F$ . With a graph attention module illustrated above, this process outputs global semantic-relations-enhanced features.

$$V_E^* = GAT(G_E), \tag{10}$$

where GAT means the graph attention module illustrated before.

This progress determines how much every pixel is affected by other pixels, where semantically more corresponding pixels may have higher attention values in the image, thus promoting the pixel-wise relations learning.

## • Attention for object-wise relations enhancement

For more refined matching of images and texts, recent regional matching methods emphasize the importance of learning the relations of objects in raw images in alignment with text phrases. VSRN [22] and ML-GCN [54] illustrated the strong potential of GCNs [48] for capturing regional relations. Different from them, this process tries to capture regional relations with a graph attention network. As seen in Fig. 3, a fully-connected graph is constructed as  $G_R = (V_R, E_R)$ , where

 $V_R$  is the regional features and  $E_R$  is the edge set defined as the affinity matrix defined in eq.(11),

$$E_R(v_R^i, v_P^j) = (v_R^i)^T v_P^j.$$
 (11)

Graph attention networks deal with the objects graph, which contains both the object features and their relations and output semantic-relations-enhanced regional representations, as shown below.

$$V_R^* = GAT(G_R). (12)$$

# D. Joint Semantic Relations Module

This part describes the kind of semantic relations that previous works lack, the object-global wise relations. As seen in Fig. 3, a multi-head graph attention module is adopted with a certain purpose to bridge the relations between regional objects and global concepts. Finally, a fusion process helps to fuse the multi-head outputs and filter out more useful information.

Firstly, the enhanced global and regional features  $V_F^*$  and  $V_R^*$  are concatenated in the object-pixel dimension into  $V_U$ ,  $V_U = \{v_u^i, ..., v_u^{n+k}\}, v_u^i \in \mathbb{R}^{D_e}$ . And a unified features graph  $G_U = (V_U, E_U)$  is obtained where  $E_U$  is the edge set defined in eq.(13).

$$E_{U}(v_{U}^{i}, v_{U}^{j}) = (v_{U}^{i})^{T} v_{U}^{j}.$$
(13)

Then, unified graph attention is conducted, just like in III-C. Different from that in III-C, here the input is the concatenated features, therefore, helping an object or a pixel learn the attention value based on all objects and pixels. With such a scheme, named joint attention, models can easily learn semantic relations between all separate elements no matter it's a regional object or a global concept. Specifically, to stabilize the learning process of self-attention, we employ multi-head attention as did in GAT [31]. As seen in Fig. 3, we feed the input  $G_U$  into K different GATs and the output is denoted as  $V_C = \{\vec{V}_C^1,...,\vec{V}_C^K\}$ . And  $\vec{V}_C^k$  is defined as below.

$$\vec{V}_C^k = Mean(GAT_k(G_U)), \tag{14}$$

where  $GAT_k$  refers to the graph attention module in the k-th JSR(joint semantic relations module) and Mean is mean-pooling.

For this module, the multi-head number K is set to  $\{1,2,4\}$ . The impact of changes in K will be discussed in Section V-B. The multi-head outputs should be fused by the fusion process.

#### • Fusion Process

With the multi-head output features  $V_C$  obtained by the joint graph attention module, we fuse them with a gated fusion layer to filter more useful information and get the final image representation.

The gated fusion layer takes two vectors  $\vec{V}_C^i$  and  $\vec{V}_C^j$  as input and outputs a fused representation.

$$\vec{V}_1 = W_1 \vec{V}_C^i, \quad \vec{V}_2 = W_2 \vec{V}_C^j, \quad t = \sigma(U_1 \vec{V}_1 + U_2 \vec{V}_2),$$

$$\vec{V} = t \odot \vec{V}_1 + (1 - t) \odot \vec{V}_2,$$
(15)

where W and U are the fully-connected layer parameters,  $\sigma$  is the sigmoid function to project coefficients to a scale 0-1.

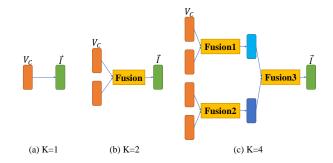


Fig. 4. Illustration of different fusion process with the change of K.

Considering the different values of K, we distinguish them in the fusion process as illustrated in Fig. 4.

- i) K = 1. No fusion is needed, the final image representation is  $\vec{I} = V_C$ .
- ii) K=2. Since  $V_C$  has two parts  $\vec{V}_C^1$  and  $\vec{V}_C^2$ , the final image representation comes from the output of one gated fusion process.

$$\vec{I} = F(\vec{V}_C^1, \vec{V}_C^2),$$
 (16)

where F is the gated fusion layer as defined in eq. (15).

*iii*) K = 4.  $V_C$  has four parts  $\vec{V}_C^1$ ,  $\vec{V}_C^2$ ,  $\vec{V}_C^3$  and  $\vec{V}_C^4$ . Totally we need three gated fusion layers.

$$\vec{I} = F_3(F_1(\vec{V}_C^1, \vec{V}_C^2), F_2(\vec{V}_C^3, \vec{V}_C^4)), \tag{17}$$

where  $F_1$ ,  $F_2$  and  $F_3$  mean the three gated fusion layers.

#### E. Learning Text Representation

Given the original sentence T corresponding to its matching image, the deep neural network embeds it into word representations. Traditionally, an RNN based network like LSTM [17] or GRU [18] is used to process the embedded word vectors, and the output hidden states are regarded as the word representations. Recently with the development of the pretraining scheme in the NLP field, another more sophisticated substitute is to use BERT [19] as the text encoder. The selfattention based transformer structures boost the representation learning of words, for the transformer structure is better at learning semantic consistency, especially when the sentence is quite long. In this paper, we adopt either GRU [18] or BERT [19] to learn word representations. Assume the maximum word number is m, so the words can be illustrated as  $W = \{w_i, ..., w_m\}, w_i \in \mathbb{R}^{D_w}$ . Then we feed them into i) the BERT-BASE encoder, which has 12 layers, and we extract the outputs of the last layer as the word representations C, ii) the GRU encoder, and the output is the representations of the words C, which is a matrix whose first dimension is the maximum words number and the second dimension is denoted as  $D_w$ . A fully-connected layer is applied to embed the features into the shared latent space where the dimension is  $D_e$ .

$$C^* = W_c C + b. (18)$$

For the text part, we construct the textual graph  $G_C = (C^*, E_C)$  where  $C^*$  serves as the nodes and  $E_C$  is defined as the edge set.

$$E_C(C_i^*, C_i^*) = (C_i^*)^T C_i^*. (19)$$

We conduct the same graph attention as in III-C to obtain finer text representation. As words in sentences always have close relations with others like "a boy **in** a red shirt", with the graph attention module, relations between nodes in the  $G_C$  are strengthened to some extend. The final text representation  $\vec{T}$  can be obtained as below.

$$\vec{T} = Mean(GAT(G_C)), \tag{20}$$

where *GAT* is the graph attention module in III-C and *Mean* refers to mean pooling on the word level.

#### F. Matching Process and Loss Function

After obtaining the two-modal representations  $\vec{I}$  and  $\vec{T}$ , a hinge-based triplet ranking loss [3] is adopted to supervise the latent space learning procedure. The loss function tries to find the hardest negatives in a mini-batch, which form the triplets with the positive ones and ground truth query. The loss function is defined below.

$$L = [\alpha + S(\vec{I}', \vec{T}) - S(\vec{I}, \vec{T})]_{+} +$$

$$[\alpha + S(\vec{I}, \vec{T}') - S(\vec{I}, \vec{T})]_{+}.$$
(21)

Here  $S(\cdot)$  refers to similarity function which is cosine similarity in our model.  $[x]_+ \equiv max(x,0)$  and  $\alpha$  is the margin.

# G. Testing Stage with Re-ranking Scheme

As did in previous works for the testing stage, the model encodes the images and texts into visual and textual feature vectors. By computing the similarities with cosine similarity, we get the similarity matrix for all the testing images and texts. Inspired by the re-ranking scheme proposed by MTFN [43], a re-ranking process reorganizes the similarity matrix to get a more accurate one. Retrieval results can be easily obtained by ranking the similarities between the query and its search items. In MTFN, as text-to-image re-ranking needs another text encoding path to compute single-modal text similarity, we merely use the image-to-text re-ranking in the experiments.

## IV. EXPERIMENTS

To evaluate our DSRAN on the image-text matching task, we perform several experiments on both image retrieval and text retrieval. Table I and Table II are the compare results with state-of-the-art methods.

## A. Datasets and Evaluation Metrics

We apply the two publicly available datasets, Microsoft COCO [32] and Flickr30K [33]. In Flickr30K, there are 31,783 images with five captions each. Following [3], the images are split into 29,000, 1000 and 1000 for training, validation and testing. As for the MSCOCO dataset, there are a total of 123,287 images, and every image has five description captions.

As did in [3], [21], [22], the splits contain 113,287 images for training, 5000 for validation and 5000 for testing. Specifically for MSCOCO, the final results are obtained either by averaging over five folds of 1k test images (referred to as 1K test set) or by directly testing the whole 5k images (referred to as 5K test set). For both image retrieval and text retrieval tasks, we record the results by calculating the recall at K (R@K) metrics defined as the proportion of the queries whose correct retrieved results are among the top-K ranking results. Specifically, we use R@1, R@5, and R@10 together with Rsum defined as below.

$$Rsum = \underbrace{R@1 + R@5 + R@10}_{\text{image retrieval}} + \underbrace{R@1 + R@5 + R@10}_{\text{text retrieval}}.$$
(22)

# B. Implementation Details

This section gives more detailed model settings and training settings for our DSRAN in the experiments. For global-wise feature maps extraction, the raw image is firstly randomly cropped and resized to  $224 \times 224$ . Moreover, we utilize the feature map of the fourth layer of Resnet152 [16] with the feature map number n set to  $7 \times 7 = 49$ . For region-wise object features extraction, we simply use the regional features given by [55], and the number of regions k is 100. Both kinds of features share the same dimension  $D_o$ , which is 2048. As for texts, we use either a pre-trained BERT-BASE [19] model, and the embedding dimension  $D_w$  is 768 or a GRU encoder [18] with a word embedding size of 300. The BERT-BASE encoder is finetuned while parameters of visual encoders ResNet152 and Faster-RCNN are fixed. The embedded latent space dimension  $D_e$  is set to 1024. The multi-head number K in the joint semantic relations module in BERT-based models (which means models with a BERT-BASE encoder) is set to 2 or 4 for Flickr30K and MSCOCO, respectively. For GRU-based models (which means models with a GRU encoder) K is set to 2 for both datasets.

For the model with BERT, experiments are performed on at least two NVIDIA 1080Ti GPU with the batch size setting to 320 for MSCOCO and 128 for Flickr30K. We train the model with an Adam optimizer [56] with a warmup rate of 0.1 for 12 and 18 epochs for Flickr30K and MSCOCO, respectively. The learning rate is set to 2*e*-5 at first and declines by ten times every 6 or 9 epochs, half of the total training epochs. While for the GRU-based model, experiments are performed on one NVIDIA 1080Ti GPU with the batch size setting to 300 for MSCOCO and 128 for Flickr30K. We train the model with an Adam optimizer for 16 and 18 epochs for Flickr30K and MSCOCO, respectively. The learning rate is set to 2*e*-4 at first and declines by ten times every 8 or 9 epochs.

#### C. Comparative Experiments with State-of-the-art Methods

We compare our DSRAN model with current state-of-the-art methods. As discussed in Section II, they are divided into three kinds, *i*) global matching methods DCCA [12], DSPE [34], MFM [44], GXN [40], VSE++ [3], MTFN [43] and TOD-Net [37], *ii*) regional matching methods SCAN [21], RDAN

TABLE I

RESULTS ON MS-COCO DATASET. METHODS ARE DIVIDED INTO THREE CATEGORIES, GLOBAL MATCHING METHODS, REGIONAL MATCHING METHODS, AND MULTI-LEVEL MATCHING METHODS THAT UNIFY THE GLOBAL AND LOCAL CONCEPTS. WE GIVE OUT BOTH PERFORMANCES ON A SINGLE MODEL OR TWO-MODELS ENSEMBLE. THE BEST RESULTS ARE IN BOLD, WHILE THE SUBOPTIMAL VALUES ARE UNDERLINED.

	Image-to-Text Text-to-Image							Im	Image-to-Text Text-to-Image					
Methods	1K Test Set							5K Test Set						
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
<b>Global Matching Method</b>	s													
DCCA	22.5	34.6	45.5	19.2	30.4	41.3	193.5	6.9	21.1	31.8	6.6	20.9	32.2	119.5
DSPE	50.1	79.7	89.2	39.6	75.2	86.9	420.7	-	-	-	-	-	-	-
MFM	58.9	86.3	92.4	47.7	81.0	90.9	457.2	-	-	-	-	-	-	-
VSE++	64.6	90.0	95.7	52.0	84.3	92.0	478.6	41.3	71.1	81.2	30.3	59.4	72.4	355.7
GXN	68.5	-	97.9	56.6	-	94.5	-	42.0	-	84.7	31.7	-	74.6	-
MTFN(re-rank)	74.3	94.9	97.9	60.1	89.1	95.0	511.3	48.3	77.6	87.3	35.9	66.1	76.1	391.3
TOD-Net(BERT-Large)	75.8	95.3	98.4	61.8	89.6	95.0	515.9	-	-	-	-	-	-	-
Regional Matching Methods														
SCAN	70.9	94.5	97.8	56.4	87.0	93.9	500.5	46.4	77.4	87.2	34.4	63.7	75.7	384.0
RDAN	74.6	96.2	98.7	61.6	89.2	94.7	515.0	-	-	-	-	-	-	-
PFAN	75.8	95.9	99.0	61.0	89.1	95.1	515.9	-	-	-	-	-	-	-
SGM	73.4	93.8	97.8	57.5	87.3	94.3	504.1	50.0	79.3	87.9	35.3	64.9	76.5	393.9
Multi-Level Matching Mo	ethods													
MDM	54.7	84.1	91.9	44.6	79.6	90.5	445.4	-	-	-	-	-	-	-
GSLS	68.9	94.1	98.0	58.6	88.2	94.9	502.7	-	-	-	-	-	-	-
CSAC	72.3	96.0	99.0	58.9	89.8	96.0	512.0	47.2	78.3	87.4	34.7	64.8	76.8	389.2
DSRAN(GRU)	76.3	94.9	98.4	62.4	89.7	95.2	516.9	51.9	81.6	89.8	39.5	70.6	81.0	414.4
DSRAN(GRU)(re-rank)	79.0	96.2	98.5	62.4	89.7	95.2	521.0	55.4	84.4	91.0	39.5	70.6	81.0	421.9
DSRAN(BERT)	77.1	95.3	98.1	62.9	89.9	95.3	518.6	53.7	82.1	89.9	40.3	$\overline{70.9}$	81.3	418.2
DSRAN(BERT)(re-rank)	<u>78.8</u>	96.1	98.5	62.9	89.9	95.3	521.5	56.3	84.2	90.7	40.3	70.9	81.3	423.7
Two-Models Ensemble														
SCAN	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3	80.4	410.9
PFAN	76.5	96.3	99.0	61.6	89.6	95.2	518.2	-	-	-	-	-	-	-
VSRN	76.2	94.8	98.2	62.8	89.7	95.1	516.8	53.0	81.1	89.4	40.5	70.6	81.1	415.7
TOD-Net(BERT-Large)	78.1	96.0	98.6	63.6	90.6	95.8	522.7	-	-	-	-	- '	-	-
DSRAN(GRU)	78.0	95.6	98.5	64.2	90.4	95.8	522.5	54.4	83.5	91.3	41.5	71.9	82.1	424.7
DSRAN(GRU)(re-rank)	80.4	96.7	98.7	64.2	90.4	95.8	526.2	57.6	85.6	91.9	41.5	71.9	82.1	430.6
DSRAN(BERT)	$\frac{33.1}{78.3}$	95.7	98.4	64.5	90.8	95.8	523.5	$\frac{57.6}{55.3}$	83.5	90.9	$\frac{11.5}{41.7}$	$\frac{72.7}{72.7}$	82.8	426.9
DSRAN(BERT)(re-rank)	80.6	96.7	98.7	64.5	90.8	95.8	527.1	57.9	85.3	92.0	41.7	72.7	82.8	432.4
	•							1 "						

[49], PFAN [46], VSRN [22] and scene-graph based method SGM [23], *iii*) multi-level matching methods GSLS [27], MDM [28], CASC [30]. Our DSRAN belongs to the multi-level matching methods. We record results from models with BERT or GRU. It should be noticed that TOD-Net uses the 24-layer BERT-Large model rather than our 12-layer BERT-BASE model. We record results whether with a re-ranking process as mentioned in Section III-G or not. Results from a single model or two-models ensemble are both recorded here. When conducting the ensemble scheme, the similarity scores from two already trained models are averaged for the final ranking process.

# • Results on MSCOCO

As shown in Table I, the highest performance of each metric is made bold, and the suboptimal values are underlined. Our DSRAN outperforms other methods, whether using an ensemble or not, except for the R@10 in the 1K test set, which may be due to noise. It is noticed that even without the re-ranking scheme, our model with either BERT or GRU still gains the best performance in most of the metrics. For the 1K test set, our model exceeds the current best TOD-Net [37] with a BERT-Large text encoder against our BERT-

BASE encoder by 3.0 and 1.0 on text retrieval and image retrieval respectively at R@1 (single model). From the table, performance gains of R@5 and R@10 are not as significant as that of R@1. This may be due to the existence of more interference sources for a given query in such a large target set. For the 5K test set, similarly, our model outnumbers the state-of-the-art VSRN [22] by 4.9 and 16.7 considering the R@1(I2T) and Rsum metric. The above outperforming proves our dual semantic relations learning scheme's effectiveness, focusing on the unified global-region visual representations learning. By applying BERT encoder, we gain a little increase in some metrics, while for others, the trend seems to be the opposite (R@1 and R@5 for single model).

#### • Results on Flickr30K

Performances on Flickr30K are shown in Table II. Our proposed DSRAN outperforms other state-of-the-art methods by a large margin. Whether with or without the re-rank scheme, our model achieves the best results in all the metrics. Compared to the previous best model VSRN [22], we increase 9.2 on text retrieval and 4.5 on image retrieval (*R*@1), with a great improvement on the Rsum metric (28.4). In case of the comparison with the current best model of the three kinds of

TABLE II

RESULTS ON FLICKR30K. THE CONFIGURATIONS ARE THE SAME AS THOSE OF MSCOCO. TOD-NET IS NO LONGER SHOWN HERE BECAUSE NO EXPERIMENTS ON THIS DATASET CAN BE FOUND IN THEIR PAPER.

Methods	Ima	age-To	-Text	Tex				
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum	
<b>Global Matching Method</b>	ls							
DCCA	27.9	56.9	68.2	26.8	52.9	66.9	299.6	
DSPE	40.3	68.9	79.9	29.7	60.1	72.1	351.0	
MFM	50.2	78.1	86.7	38.2	70.1	80.2	403.5	
VSE++	52.9	80.5	87.2	39.6	70.1	79.5	409.8	
GXN	56.8	-	89.6	41.5	-	80.1	-	
MTFN	65.3	88.3	93.3	52.0	80.1	86.1	465.1	
Regional Matching Meth	ods							
SCAN	67.9	89.0	94.4	43.9	74.2	82.8	452.2	
RDAN	68.1	91.0	95.9	54.1	80.9	87.2	477.2	
PFAN	67.6	90.0	93.8	45.7	74.7	83.6	455.4	
SGM	71.8	91.7	95.5	53.5	79.6	86.5	478.6	
Multi-Level Matching Me	ethods	1						
MDM	44.9	75.4	84.4	34.4	67.0	77.7	384.0	
GSLS	68.2	89.1	94.5	43.4	73.5	82.5	451.2	
CASC	68.5	90.6	95.9	50.2	78.3	86.3	469.8	
DSRAN(GRU)	72.6	93.6	96.3	<u>56.3</u>	84.0	89.8	492.6	
DSRAN(GRU)(re-rank)	<u>75.7</u>	94.7	96.8	<u>56.3</u>	84.0	89.8	497.3	
DSRAN(BERT)	75.3	94.4	97.6	57.3	84.8	90.9	500.3	
DSRAN(BERT)(re-rank)	78.6	95.6	97.6	57.3	84.8	90.9	504.8	
Two-Models Ensemble								
SCAN	67.4	90.3	95.8	48.6	77.7	85.2	465.0	
PFAN	70.0	91.8	95.0	50.4	78.7	86.1	472.0	
VSRN	71.3	90.6	96.0	54.7	81.8	88.2	482.6	
DSRAN(GRU)	74.9	94.5	97.0	<u>58.6</u>	85.8	91.3	502.1	
DSRAN(GRU)(re-rank)	79.6	95.6	97.5	58.6	85.8	91.3	508.4	
DSRAN(BERT)	77.8	95.1	<u>97.6</u>	59.2	86.0	91.9	507.6	
DSRAN(BERT)(re-rank)	80.5	<u>95.5</u>	97.9	59.2	86.0	91.9	511.0	

methods, i) for MTFN, our model outperforms it by 13.3 and 5.3 considering R@1, ii) for SGM, the performance gain in R@1 is 6.8 and 3.8 respectively for text retrieval and image retrieval, iii) for CASC, our R@1 outnumbers theirs by 10.1 and 7.1. The I2T performance gain seems more obvious than that of T2I performance gain, which is the result of conducting only I2T re-ranking, as stated in III-G. At the same time, the performances increase seems more salient when using BERT encoder for texts, which may be due to the powerful capabilities in processing longer sentences of BERT.

# · Results Analysis

From the tables, we get some more detailed findings.

- a) Our model is superior to all the global matching methods in all the metrics. Since our model takes both global and local regional features into account, global matching methods that only focus on the global concept alignment seem insufficient for refined image-text matching. Further, the performance gain against regional matching methods, which only concentrate on region-word alignment, proves that the combination of global and regional alignment benefits the matching as did in our method.
- b) The proposed DSRAN also outperforms previous multilevel matching methods. Although MDM, GSLS, and CASC integrate both local region-word matching and global semantics matching, they lack the process of learning relations

between regions and the relations between regional objects and global concepts.

c) Compared with VSRN, who limitedly considers learning relations between regional objects in the images, our proposed global-region relations learning boosts this task. There are always relations between salient objects and hard-to-catch global concepts in the captions. Thus our dual semantic relations learning is significant for accurate matching.

#### D. Visualization of Retrieval Results

To validate our dual semantic relations learning scheme's practical matching results, we list the top-3 ranking items of the image query and the top-4 ranking items of the text query. We use Flickr30K dataset and the model with the highest Rsum without re-ranking to do this experiment (that is DSRAN(BERT) in Table II). Specifically, in Fig. 5, we list results from two models, i) VSRN [22] which only considers the regional relations, ii) our DSRAN which integrates both regional relations and region-global relations. Here I2T means text retrieval with the image query, and for T2I, vice versa. On the illustration's far-right, we highlight the sentences that VSRN retrieves wrong, but DSRAN retrieves right. Specially, we mark the objects and global contexts together with their relations with different colors. Moreover, the relation shares the same color with the object or global concept following it. We additionally underline the global concepts. For example, for the sentence on the second row of the table "A little girl wearing a black swimming suit holding a shovel at a beach", "girl", "black swimming suit", "shovel" are the salient objects and "beach" is a global concept while "wearing", "holding", "at" are the relations between them.

From Fig. 5, it is obvious that a decisive factor that VSRN fails to match the image-text pair is that their model lacks the learning of region-global relations. As seen in the second sample, although the first retrieval caption for the query image still contains a key object with relation to the image region which is "smiling young girl". However, except for the salient object, global concepts and their relations with the object are equally significant. For the image, the global concept is "beach" and the relation is "at". Simply ignoring this kind of relations may lead to image-text mismatch. For example, in line one, the mismatch of VSRN comes from "boy in the air" and in row three, the mismatch comes from "sunlight reflecting the water", "men standing against water". And in row four, the mismatch still comes from "at a playground" and "by the ocean".

#### V. ABLATION STUDY AND ANALYSIS

In this section, firstly in V-A we do several ablation studies considering the dual semantic relations enhancement schemes used in our model, *i*) separate semantic relations module(SSR), *ii*) joint semantic relations module(JSR) and the integration of dual-paths including the regional path and the global path. Then in V-B, we analyze the graph attention modules used in both SSR and JSR. At last in V-C, we analyze the training process of our model and the effectiveness of batch normalization.

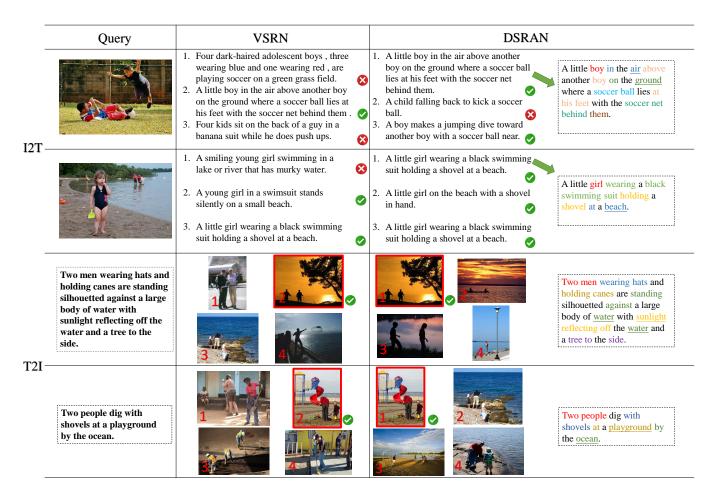


Fig. 5. The retrieval results of our DSRAN and VSRN, which only considers regional relations. Experiments are conducted using the Flickr30K dataset and on the models with the highest *Rsum*. Two instances are shown for I2T retrieval and T2I retrieval each. We give the top-three ranking texts for every image query and top-four ranking images for every text query. In the figure's right column, we highlight the ground-truth caption by marking objects, global concepts, and relations with several colors. Global concepts are underlined.

## A. Ablation Studies on DSRAN

For the visual part of our DSRAN, there are two paths, one for global features extraction and another for local regional features extraction. Thus, for ablation studies, we conduct experiments using a) only global path, b) only regional path, and c) both two paths. There are two main semantic relations modules in our DSRAN, the separate semantic relations module (referred to as SSR) and the joint semantic relations module (referred to as JSR). We perform ablation experiments on Flickr30K [33] test set with or without the modules. Thus, there are total 8 experiments for BERT-based models and another 8 for GRU-based models as listed in Table III. It should be noticed that no re-ranking is used in the experiments. In the table, "Global" refers to using the global path, and "Regional" refers to using the regional path. "SSR" and "JSR" refer to whether using those two semantic relations modules. "BERT" and "GRU" refer to the text encoder use in the model. For example, in line 6, "Regional", "Global", "SSR" and "BERT" are chosen, which means we use both the global and regional paths with only the separate semantic relations module and fuse them with the gated fusion layer based on BERT encoder. For the first four lines numbered 1-4 and 9-12, these are experiments using only one path. Thus no JSR is

added for these four. For the last four lines numbered 5-8 and 13-16, these are experiments using both paths with or without these two semantic relations modules. It is noticed that the configurations of numbers 8 and 16 are the same as our final models for comparison.

As shown in Table III, for lines 1-4 and 9-12, which simply use a single path for image representation learning, single regional object-word alignment is superior to single global semantics alignment. Moreover, the use of SSR boosts the performances in a significant way proving the effectiveness of our proposed SSR aiming at learning whether pixel-wise global relations or object-wise regional relations. When using both paths in line 5-8 and 13-16, not only can we see that with our proposed dual semantic relations learning scheme, images and texts are better matched in a fine-grained way, considering both regional relations and region-global relations, but also this kind of global-regional integration scheme is superior to single-path visual encoding while computing only one similarity in the end. Comparing models with different text encoders (BERT or GRU), although GRU-based models perform lower than BERT-based models, the performance gains from the proposed modules seem more significant for the former.

#### TABLE III

ABLATION STUDIES ON DIFFERENT MODEL SETTINGS. "REGIONAL" AND "GLOBAL" REFER TO USING WHETHER A SINGLE-PATH MODEL OR BOTH. "SSR" REFERS TO THE SEPARATE SEMANTIC RELATIONS MODULE WHILE "JSR" IS THE JOINT SEMANTIC RELATIONS MODULE. "BERT" AND "GRU" DETERMINE THE TEXT ENCODER. THE TOP-8 ROWS ARE MODELS WITH BERT ENCODER WHILE THE REST ARE THOSE WITH GRU ENCODER.

The highest value of each metric is made bold. We run this ablation study on the Flickr30K dataset. We record results from single models without the re-ranking process.

Number		I	Model Se	ttings			Image-to-Text			Text-to-Image			
	Regional	Global	SRR	JRR	BERT	GRU	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
1		✓			<b>√</b>		48.1	78.1	87.2	36.4	68.1	78.6	396.4
2		✓	✓		$\checkmark$		52.9	79.8	88.5	38.2	69.7	79.9	409.0
3	✓				✓		58.4	84.8	91.9	44.5	75.3	84.2	439.1
4	✓		✓		$\checkmark$		73.2	92.3	96.4	56.3	83.6	89.8	491.6
5	✓	✓			$\checkmark$		68.0	91.7	96.0	51.6	81.2	88.4	476.9
6	✓	✓	✓		$\checkmark$		72.8	93.4	97.1	57.1	84.5	90.8	495.6
7	✓	✓		$\checkmark$	$\checkmark$		73.4	92.9	96.9	57.2	84.7	90.8	495.9
8	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$		75.3	94.4	97.6	57.3	84.8	90.9	500.3
9		<b>√</b>				✓	43.3	73.9	83.4	33.7	63.9	74.7	372.9
10		✓	✓			✓	52.6	78.5	85.4	39.2	69.4	79.1	404.3
11	✓					✓	61.2	87.1	93.2	45.0	74.7	83.5	444.7
12	✓		✓			✓	67.7	89.4	94.8	51.4	79.8	87.9	470.9
13	✓	$\checkmark$				✓	66.4	89.4	95.1	48.9	78.1	86.3	464.2
14	✓	$\checkmark$	$\checkmark$			✓	69.3	90.8	96.3	54.5	81.7	89.2	481.8
15	✓	$\checkmark$		✓		✓	70.1	90.7	95.5	53.2	81.8	88.9	480.2
16	✓	$\checkmark$	$\checkmark$	$\checkmark$		$\checkmark$	72.6	93.6	96.3	56.3	84.0	89.8	492.6

# B. Analysis on Graph Attention Module

In both semantic relations modules, we make a little modification to traditional GATs [31] to suit our model and apply them to enhance relations learning. The model successfully learns the relations-enhanced features by constructing two separate fully-connected graphs for global features and regional features. Here we analyze GATs used in SSR and JSR, respectively.

1) GATs in separate semantic relations module: As seen in lines 1-4 and 9-12 in Table III, a noticeable performance gain can be obtained with SSR for both single-path methods. Concretely, with SSR, Rsum of the single global path increases from 396.4 to 409.0. The performance gain comes from learning global pixel-wise semantic relations. While Rsum of the single regional path has much more significant growth of 52.5 from 439.1. This comes from the enhancement of regional relations learning realized by our graph attention module.

In Section III-C, we construct a global visual graph  $G_F$ and regional visual graph  $G_R$  and utilize the GATs to enhance global pixel-wise relations and regional object-wise relations. Specifically, for the image feature map directly extracted by CNNs, it has a grid-like structure, and the arrangement is relatively uniform. While GATs are better at processing graphstructured data, such as the image regions with irregular distribution as shown in Fig. 3. Further, to illustrate the effectiveness of the GATs on capturing semantic relations, we visualize the attention using Flickr30K dataset. Considering that the final image representation should pay more attention to the salient objects in the raw images. Firstly we compute the similarities of the final image representation  $\vec{I}$  and CNN-extracted global features  $V_F$  or Faster-RCNN-extracted regional features  $V_R$ with simply dot production. Thus, every area or every region has a similarity value with the final visual representation. Then we rank the values, thus the areas or regions with higher ranks are marked brighter in the attention maps, as shown in Fig. 6. The left picture is the original image, the middle one is the image with salient regions highlighted, and the right one is the picture with some areas marked brighter. We choose the top-50 regions ranked by similarity score for regional attention visualization whicle for global attention, the raw image is cropped into  $7 \times 7 = 49$  areas. Experiments are conducted whether using model 2 or model 4, as in Table III. Four instances are given.

Focusing on the middle column, the regional attention pictures, the final image representation successfully concentrates on the images' salient objects. However, it is hard for global attention pictures to pay attention to the salient areas concerning the words in sentences. To sum up, as discussed in GAT [31], what is more powerful about graph attention network is that it deals with the data of an irregular graph structure. Thus, when dealing with the regional visual graph, GATs help stress more salient nodes, which contribute more to the alignment with words in sentences. As for the global attention, although GATs can not pay attention to more significant areas, the relationsenhanced global features can still be used as a supplement to the full image information, and promote the later joint relations learning.

2) GATs in joint semantic relations module: In Section III-D, we learn the joint semantic relations by dint of GATs. With the strong ability of GATs in capturing node-level relations, the JSR successfully learns the region-global relations. As discussed before, for more profound attention in the graph attention module, we apply multi-head graph attention, and for different multi-head number K, the fusion process varies. As discussed in [53], the use of multi-head attention enables each head to focus on different priorities, so more regional-global relations can be captured by the model. Fig. 7 records the Rsums with the change of K on three different testing sets Flickr30K, COCO-1k, and COCO-5k using (a) BERT

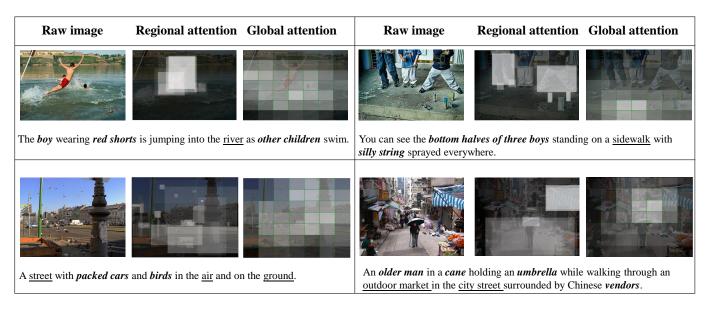


Fig. 6. Visualization of regional and global attention achieved by GATs using Flickr30K dataset. The left column is the raw image, and the others are regional and global attention pictures. We have the salient objects in the texts in bold, and italics and underline global context like "river", "ground" and "street". In the attention pictures, more critical areas are highlighted brighter. We show the results of pictures with a brief background on the first row while pictures on the second row have more complex backgrounds.

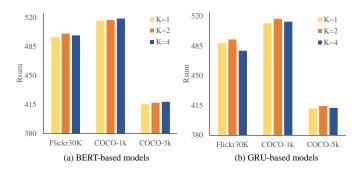


Fig. 7. How Rsum goes as K grows for different test sets considering (a) models with BERT and (b) models with GRU.

and (b) GRU. For Flickr30K, when K equals 2, Rsum is the highest while for COCO-1k and COCO-5k, the top comes when K is 4 for BERT and 2 for GRU. The difference between Flickr30K and COCO is that COCO has a richer training set, so even if the network gets deeper, the model can still learn more abundant relations. However, when applying GRU, the deficiency of masked pre-training limits the performance when K gets larger.

# C. Analysis on Training Process

1) Analysis of training epochs: In this part, to reduce the influence of underfitting and overfitting, we aim to find out the best training epochs for Flickr30K and MSCOCO, respectively. We conduct experiments using the full models whose configuration is the same as number 8 and 16 in Table. III. It should be noticed that for each experiment, the learning rate declines by ten times every half of the total epochs. As seen, for Flickr30K test set, the curve peaks when the epoch is set to 12 for BERT and 16 for GRU. Moreover,

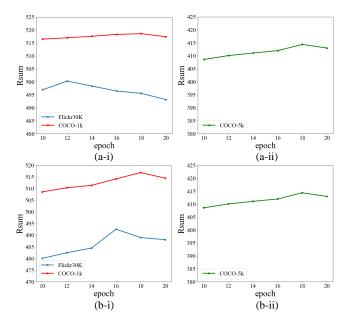


Fig. 8. Rsum results for models with different training epochs on a) BERT-based models, b) GRU-based models, and i) Flickr30K and COCO-1k test set, ii) COCO-5k test set.

for MSCOCO, as seen in both 1k and 5k test sets, the training epoch should be set to 18.

2) Analysis on batch normalization: When building the graph attention module in Section III-C, we add a batch normalization at the end of the module. As discussed in [57], a BN layer can prevent the gradient explosion or dispersion, improve the robustness of the model to different super parameters (learning rate, initialization), and make most of the activation functions far away from its saturation region. All these properties of BN can help us to train the network

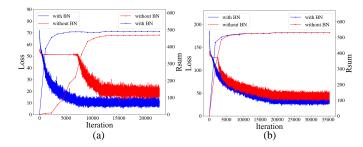


Fig. 9. The plots of training losses and Rsum testing on a) Flickr30K validation set, b) COCO validation set.

quickly and robustly. Furthermore, the critical point is that the BN changes the optimization problem and makes the optimization space very smooth. When training our model, the batch normalization layer benefits the training process. As shown in Fig. 9, we record the process of loss decline and the *Rsum* on the validation set with or without a BN layer on BERT-based models. The use of batch normalization makes the optimization more smooth, and the loss drops faster. What's more, it promotes the training process to converge earlier and boosts the final evaluation performances.

#### VI. CONCLUSION

In this paper, we focus on the visual semantic relations learning for enhanced image-text matching. Further, a dual semantic relations attention network (DSRAN) with different kinds of relations modules applied to capture both the objectlevel semantic relations and global-regional semantic relations. The learned dual-relations-enhanced visual representations can better match their textual counterparts whose words are inherently related in both object level and global-region level, thus promoting the matching procedure. Quantitative experiments show our model's successful target-oriented designs, and such a model outperforms previous methods on the image-text matching task on the two widely used datasets MSCOCO and Flickr30K. Further, we do ablation studies proving the effectiveness of the two main modules targeting dual semantic relations learning. In the future, we are looking forward to applying this kind of dual semantic relations learning to more cross-modal tasks such as image captioning and visual question answering.

# REFERENCES

- D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.
- [2] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM international* conference on Multimedia, 2017, pp. 154–162.
- [3] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," arXiv preprint arXiv:1707.05612, 2017.
- [4] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Transactions on circuits and systems for video technology*, vol. 28, no. 9, pp. 2372–2385, 2017.
- [5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

- [6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on* computer vision and pattern recognition, 2018, pp. 6077–6086.
- [7] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, "Learning to evaluate image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5804–5812.
- [8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference* on computer vision and pattern recognition, 2015, pp. 3156–3164.
- [9] X. Huang, Y. Peng, and Z. Wen, "Visual-textual hybrid sequence matching for joint reasoning," *IEEE Transactions on Cybernetics*, 2020.
- [10] X. He and Y. Peng, "Fine-grained visual-textual representation learning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 520–531, 2019.
- [11] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference* on computer vision and pattern recognition, 2015, pp. 3128–3137.
- [12] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, 2013, pp. 1247–1255.
- [13] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," arXiv preprint arXiv:1411.2539, 2014.
- [14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural infor*mation processing systems, 2012, pp. 1097–1105.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision* and pattern recognition, 2016, pp. 770–778.
- [17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
- [20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural* information processing systems, 2015, pp. 91–99.
- [21] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference* on Computer Vision (ECCV), 2018, pp. 201–216.
- [22] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4654–4662.
- [23] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1508–1517.
- [24] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5831–5840.
- [25] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proceedings of the fourth workshop* on vision and language, 2015, pp. 70–80.
- [26] J. Qi, Y. Peng, and Y. Yuan, "Cross-media multi-level alignment with relation attention network," arXiv preprint arXiv:1804.09539, 2018.
- [27] Z. Li, F. Ling, C. Zhang, and H. Ma, "Combining global and local similarity for cross-media retrieval," *IEEE Access*, vol. 8, pp. 21847– 21856, 2020.
- [28] L. Ma, W. Jiang, Z. Jie, and X. Wang, "Bidirectional image-sentence retrieval by local and global deep matching," *Neurocomputing*, vol. 345, pp. 36–44, 2019.
- [29] F. Huang, X. Zhang, Z. Zhao, and Z. Li, "Bi-directional spatial-semantic attention networks for image-text matching," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 2008–2020, 2018.
- [30] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image-text matching," IEEE Transactions on Neural Networks and Learning Systems, 2020.

- [31] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," arXiv preprint arXiv:1710.10903, 2017.
- [32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [33] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.
- [34] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE conference on* computer vision and pattern recognition, 2016, pp. 5005–5013.
- [35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.
- [36] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.
- [37] T. Matsubara, "Target-oriented deformation of visual-semantic embedding space," arXiv preprint arXiv:1910.06514, 2019.
- [38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672– 2680.
- [39] Y. Peng and J. Qi, "Cm-gans: Cross-modal generative adversarial networks for common representation learning," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 15, no. 1, pp. 1–24, 2019.
- [40] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition, 2018, pp. 7181–7189.
- [41] X. Wen, Z. Han, and Y.-S. Liu, "Cmpd: Using cross memory network with pair discrimination for image-text retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.
- [42] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 299–307.
- [43] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, "Matching images and text with multi-modal tensor fusion and reranking," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 12–20.
- [44] L. Ma, W. Jiang, Z. Jie, Y.-G. Jiang, and W. Liu, "Matching image and sentence with multi-faceted representations," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
- [45] Z. Ji, H. Wang, J. Han, and Y. Pang, "Saliency-guided attention network for image-sentence matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5754–5763.
- [46] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, and X. Fan, "Position focused attention network for image-text matching," arXiv preprint arXiv:1907.09748, 2019.
- [47] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017
- [48] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.
- [49] Z. Hu, Y. Luo, J. Lin, Y. Yan, and J. Chen, "Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching." in *IJCAI*, 2019, pp. 789–795.
- [50] Y. Peng and J. Chi, "Unsupervised cross-media retrieval using domain adaptation with scene graph," *IEEE Transactions on Circuits and Sys*tems for Video Technology, 2019.
- [51] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.
- [52] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10313–10322.
- [53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [54] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the*

- IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5177–5186.
- [55] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa." in AAAI, 2020, pp. 13 041–13 049.
- [56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
- [57] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.