Dual Semantic Relationship Attention Network for Image-Text Matching

Keyu Wen
Department of Electronic Engineering
Fudan University
Shanghai 200433, China
kywen19@fudan.edu.cn

Xiaodong Gu
Department of Electronic Engineering
Fudan University
Shanghai 200433, China
xdgu@fudan.edu.cn

***Abstract*—Image-Text Matching is one major task in cross-modal information processing. The main challenge is to learn the unified vision and language representations. Previous methods that perform well on this task primarily focus on the region features in images corresponding to the words in sentences. However, this will cause the regional features to lose contact with the global context, leading to the mismatch with those non-object words in some sentences. In this work, in order to alleviate this problem, a novel *Dual Semantic Relationship Attention Network* is proposed which mainly consists of two modules, separate semantic relationship module and the joint semantic relationship module. With these two modules, different hierarchies of semantic relationships are learned simultaneously, thus promoting the image-text matching process. Quantitative experiments have been performed on MS-COCO and Flickr-30K and our method outperforms previous approaches by a large margin due to the effectiveness of the dual semantic relationship attention scheme.**

***Index Terms*—cross-modal, retrieval, attention, semantic relationship**

I. Introduction

Different from traditional single-modal retrieval, image-text matching [31] 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 with the query text.

Recently with the development of deep neural networks, latent space learning methods [1] [2] [3] stand as a fundamental solution to this task. Traditionally image and text inputs are separated encoded by convolution-based networks [4] [5] [6] and RNN-based networks like LSTM [7] or GRU [8], after which a similarity function is used to measure the distance between two-modal representations. More recently, text representations can as well obtained with pre-trained transformer-based models like BERT [9], which are comparable to the pre-trained CNNs in the image channel. At last, a triplet-based ranking loss function [2] supervises the training and the best unified latent space is learned.

A more refined way is to extract the local regional features using faster R-CNN [12], which is called bottom-up attention [13] 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 [6]. SCAN [14] firstly introduced this scheme into image-text matching task. VSRN [15] took a step further to learn the relationship between objects in the raw image using graph convolutional network. GSLS [36] boosts the combination of image and text information by extracting both local and global features of images and captions and learning their similarities simultaneously. Although quite successful, these methods lack the emphasis on relationship between objects and non-object 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.

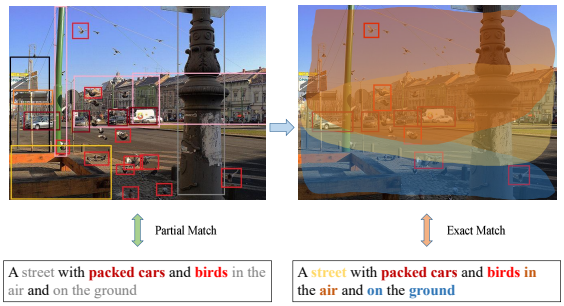

Fig. 1. The proposed DSRAN learns semantic relationship between regional objects as well as objects and global context. With only regional features, the visual representations fail to match the corresponding words and relationship like "birds **in** the air", "birds **on** the road" or "street **with** cars".

Thus, based on previous work, this paper proposes a Dual Semantic Relationship Attention Network(DSRAN) to address this problem. Intuitively there are two main modules in the DSRAN, the separate semantic relationship module and the joint semantic relationship module. The separate semantic relationship module is designed for capturing both objects and their semantic relationship. Specifically, because of the efficiency and effectiveness of GATs [35] when learning the nodes relationship, the module uses two separate graph attention networks to learn pixel-wise semantic relationship and regional relationship at the same time. The second module,

This work was supported in part by National Natural Science Foundation of China under grants 61771145 and 61371148.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

joint semantic relationship module, aims to find the semantic relationship between local objects and global pixel-wise concepts. A unified graph attention network is used to achieve this. After these two principal relationship-oriented modules, the similarity scores of the obtained image features and text features can be calculated and further update the network parameters with the loss function as previous works did. Details will be discussed later.

To verify the validity of our proposed model, we test our model on both MSCOCO [16] and Flickr30K [17] datasets. Experimental results show that our model outperforms the currently state-of-the-art method on both datasets which prove the effectiveness of our design.

Our contributions are summarized as below.

(a) We propose a novel Dual Semantic Relationship Attention Network(DSRAN) in order to strengthen the relationship between regional objects and global concepts in the learned visual representations while considering the relationship among objects themselves.

(b) The proposed DSRAN outperforms previous works on the image-text matching task. Specifically, on MSCOCO our model outnumbers the current best model VSRN [15] by 2.5% for image retrieval and 3.8% for text retrieval (Recall@1 using 5K test set). And on Flickr30K, the increase is more significant which is 9.7% for image retrieval and 7.0% for text retrieval (Recall@1).

II. Related Work

A. Image-Text Matching

The Image-Text Matching task can be regarded as one of the most fundamental tasks in cross-modal retrieval. In this task, previous works mainly focus on the latent space learning proposed by CCA [33] which mainly relies on linear projection. Further, with the development of deep learning, Visual Semantic Embedding [1] uses deep neural networks to project visual and text features into the latent space separately. [2] proposed to use a hard-negative based triplet ranking loss instead of the traditional cross-entropy loss to supervise the training which is followed by most recent works. The visual features are basically extracted by convolution-based models like [4] [6] pre-trained on ImageNet [24] until SCAN [14] introduces the bottom-up attention scheme [13] and used Faster-RCNN [12] pre-trained on Visual Genome [18] to extract more semantic object-level visual features. Further, VSRN [15] applies graph convolutional network [19] to conduct visual-reasoning corresponding to the word-level semantic meanings in texts. For text modality, traditionally original words are embedded into word vectors and fed into an RNN-based encoder like LSTM [7] or GRU [8]. With the success of pretraining in NLP field like BERT [9], GPT [10], and XLNet [11], more specific words representations can be learned which are comparable with their visual counterparts.

B. Attention Mechanism

Our work applies different kinds of attention mechanisms to handle various relationship problems. Self-attention mechanism [25] has brought the NLP field into a brand-new era. Learning from its success in text modality, in CV field [20] applies self-attention to image generation. Word embeddings or visual feature maps play as three roles as query, key, value, after which semantically more important words or regions are paid more attention. In addition to this, SCAN [14] proposed stacked cross attention between image and text modalities to learn the correspondence between words and objects. To handle topology data, in GAT [35], self attention and relational modules are combined on the graph structures for nodes classification.

C. Visual Relationship Learning

In the cross-modal field, to better match visual and text modalities, researchers pay more attention to the visual relationship learning because they believe it's of significance for machines to learn not only the objects but also their relationship just as the importance of learning relationship between words in natural language processing. In [34], a multi-label CNN extracts regional semantic concepts and learn their order relationship. The propose of scene graph, which produces a knowledge graph based on the objects and their relationship in the raw image enhances connections with the text modality [21] [22] [23]. VSRN [15] tries to capture the relationship between separated objects in the image using graph convolutional networks [19]. These works successfully learn the object-level relationship while ignoring the relationship between objects and global elements such as the atmosphere, the environment, the surroundings, and the background, which is emphasized and resolved by our method.

III. Proposed Method

In this section, we detail our proposed Dual Semantic Relationship Attention Network(DSRAN). As shown is Fig. 2, 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-A). Two modules are followed, the first of which is the separate semantic relationship module aiming to learn the object-level semantic relationship (III-B). The second is the joint semantic relationship module which is designed for capturing relationship across objects and global concepts (III-C). For the text part, a pre-trained BERT-base model [9] extracts the words representations corresponding to the image features (III-D). In the end, with the cross-modal representations, we can calculate the similarity scores and update the network parameters with the loss function(III-E).

A. 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 [6] pretrained on ImageNet [24] 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

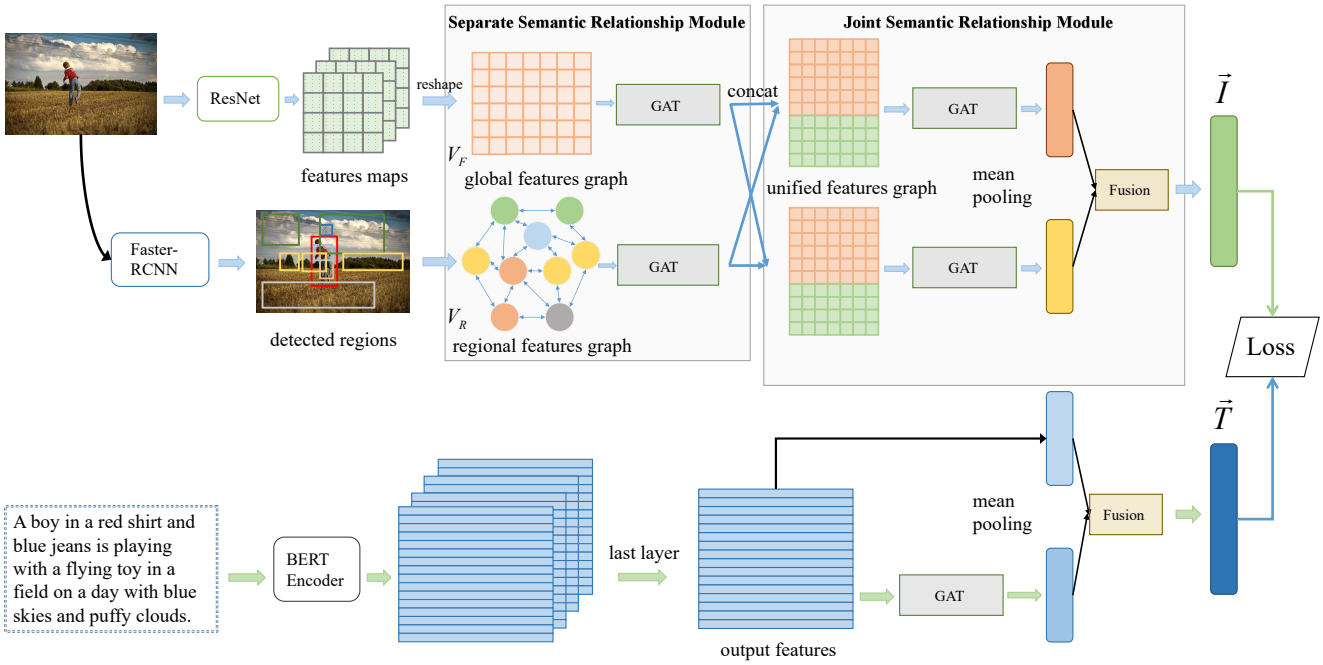

Fig. 2. An overview of the proposed Dual Semantic Relationship Attention Network. Two semantic relationship learning modules are applied to enhance both object-level relations and unified global-region relations. A self-attention mechanism and a graph convolution separately the former while two different attention mechanisms are used for the latter. The caption is fed to a pre-trained BERT-base encoder and we take the last layer of the output features. And a residual structure is designed for text representation learning.

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 [13], the objects are firstly detected by a Faster-RCNN [12] pretrained on Visual-Genome [18] and then fed into a backbone Resnet101 and the final 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.

$$V_F = W_f F + b_f \tag{1}$$

$$V_R = W_r R + b_r \tag{2}$$

W_f and W_r are the weight matrixes 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.

B. Separate Semantic Relationship Module

Targeting at dual-levels of features, we design two separate semantic relationship enhancement models to learn the enhanced pixel-wise relationship and object-wise relationship. 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 [35], we compute attention coefficients and normalize them with softmax function.

$$e_{ij} = a(W_q v_i, W_k v_j) \tag{3}$$

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

W_q and W_k are learnable parameters. In case of memory explosion, different from using the feed-forward neural network, we simply compute the attention coefficients with multi-head dot production [25].

$$a(W_q v_i, W_k v_j) = W_q v_i (W_k v_j)^T / \sqrt{d} \tag{5}$$

In this paper, the multihead num is set to 8 thus d equals to $D/8$. Thus, with a nonlinear activation function, the final output feature can be computed.

$$v_i^{'} = ReLU(\sum_{j \in N_i} \alpha_{ij} W_v v_j) \tag{6}$$

Same as [35], we employ multi-head attention.

$$v_i^{'} = \|_{k=1}^{K} ReLU(\sum_{j \in N_i} \alpha_{ij}^k W_v^k v_j) \tag{7}$$

In eq.(7) $\|$ means concatenation and K is the multi-head num. A batch normalization is followed. Here we finish the construction of a graph attention module.

- Attention for pixel-wise relationship enhancement

Obtaining the global features V_F, firstly we construct the global visual graph $G_F = (V_F, E_F)$. With a graph attention

module illustrated above, this process outputs global semantic-relationship-enhanced features.

$$V_F^* = GAT(G_F) \tag{8}$$

GAT means the graph attention module illustrated before. K is set to 1. This attention proceeding can be repeated for x times for deeper attention.

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

- Attention for object-wise relationship enhancement

VSRN [15] and ML-GCN [32] illustrate the strong potential of GCNs [19] for capturing regional relationship. Different from them, this process tried to capture the regional relations with a graph attention network. As seen in Fig. 2, a fully-connected graph is constructed as $G_R = (V_R, E_R)$, where V_R is the regional features. Graph Attention networks deal with the objects graph which contain both the object features and their relationship and output semantic-relationship-enhanced regional representations, as shown below.

$$V_R^* = GAT(G_R) \tag{9}$$

Samely K is set to 1 and the GAT can be repeated for x times for better representation learning.

C. Joint Semantic Relationship Module

In this part, we describe the kind of semantic relationship that previous works lack, the object-global wise relationship. As seen in Fig. 2, 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 fuses the multi-head outputs.

Firstly, the enhanced global and region features V_F^* and V_R^* are concatenated in the object-pixel dimension into V_U. And a unified features graph $G_U = (V_U, E_U)$ is obtained. Then, a unified graph attention is conducted just like in III-B. Different from that in III-B, 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 relationship between all separate elements no matter it's a regional object or a global concept.

$$V_C = GAT(G_U) \tag{10}$$

Here K is set to 2. The concatenated multi-head outputs should be fused.

- Fusion Process

With the multi-head output features V_C obtained by joint graph attention module, we fuse them with simply a fully-connected layer to get the final image representation. Firstly by mean pooling, matrixes $V_C = \{v_c^i, ..., v_c^n\}, v_c^i \in \mathbb{R}^{2D_e}$ transform into vectors, after which they are concatenated and fed into a fully-connected layer.

$$\vec{V} = Mean(V_C) \tag{11}$$

$$\vec{I} = W_I\vec{V} + b_I \tag{12}$$

Here $Mean$ is mean-pooling, W_I and b_I are the fully-connected layer parameters.

D. Getting Text Representation With Bert

Given the original sentence T corresponding to its matching image, the deep neural network embeds it into words representations. Traditionally, an RNN based network like LSTM [7] or GRU [8] is used to process the embedded word vectors and the output hidden states are regarded as the words representations. Recently with the development of the pre-training scheme in NLP filed, another more sophisticated substitute is to use BERT [9] as the text encoder. The self-attention based transformer structures boost the words representations learning and match the attention mechanisms used on the image side. Assume the maximum word num 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 the BERT-base encoder which has 12 layers and we extract the outputs of the last layer as the word representations C which is a matrix whose first dimension is the maximum words num and the second dimension is noted as D_w. A fully-connect layer is applied to embed the features into the shared latent space where the dimension is D_e.

$$C^* = W_c C + b \tag{13}$$

To match the dimension of final image representation $\vec{I}$, we conduct the same graph attention process and the fusion process exactly the same way in III-B and III-C. The final text representation $\vec{T}$ can be obtained as below.

$$\vec{T} = F(Concat(Mean(C^*), Mean(GAT(C^*)))) \tag{14}$$

F refers to the fusion process in III-C, SA refers to graph attention module in III-B and $Mean$ refers to mean pooling on the word level. K of GAT is set to 1.

E. Matching Process and Loss Function

After obtaining the two-modality representations $\vec{I}$ and $\vec{T}$, a hinge-based triplet ranking loss [2] 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 groundtruth query. The loss function is defined as below.

$$\begin{aligned} L = &[\alpha + S(\vec{I}', \vec{T}) - S(\vec{I}, \vec{T})]_+ + \\ &[\alpha + S(\vec{I}, \vec{T}') - S(\vec{I}, \vec{T})]_+ \end{aligned} \tag{15}$$

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

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.

TABLE I

RESULTS ON MS-COCO DATASET. METHODS ARE DIVIDED INTO THREE CATEGORIES, GLOBAL-WISE, REGION-WISE AND OUR GLOBAL-REGION UNIFIED KIND. WE GIVE OUT BOTH PERFORMANCES ON A SINGLE MODEL OR TWO-MODELS ENSEMBLE. THE BEST RESULTS ARE IN BOLD.

Methods	**Image-to-Text**			**Text-to-Image**				**Image-to-Text**			**Text-to-Image**			
	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
Global-Wise Visual Representations														
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
MTFN	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	-	-	-	-	-	-	-
Region-Wise Visual Representations														
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
PFAN	75.8	95.9	**99.0**	61.0	89.1	95.1	515.9	-	-	-	-	-	-	-
Global-Region Unified Visual Representations(Ours)														
DSRAN	**76.5**	**96.0**	98.4	**62.7**	**89.7**	**95.2**	**518.4**	**52.7**	**81.5**	**90.3**	**39.9**	**70.9**	**81.1**	**416.4**
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	**78.2**	**96.3**	98.6	**64.2**	**90.6**	**95.8**	**523.7**	**55.0**	**83.0**	**90.6**	**41.5**	**72.2**	**82.4**	**424.7**

TABLE II

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

Methods	**Image-To-Text**			**Text-To-Image**			
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
Global-Wise Visual Representations							
VSE++	52.9	80.5	87.2	39.6	70.1	79.5	409.8
MTFN	65.3	88.3	93.3	52.0	80.1	86.1	465.1
Region-Wise Visual Representations							
SCAN	67.9	89.0	94.4	43.9	74.2	82.8	452.2
PFAN	67.6	90.0	93.8	45.7	74.7	83.6	455.4
Global-Region Unified Visual Representations(Ours)							
DSRAN	**74.6**	**93.8**	**97.2**	**57.8**	**84.8**	**90.6**	**498.7**
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	**76.3**	**94.4**	**97.5**	**60.0**	**85.8**	**91.9**	**505.9**

A. Datasets and Evaluation Metrics

We apply the two publicly available Microsoft COCO dataset [16] and Flickr30K dataset [17]. In Flickr30K, there are 31,783 images with 5 captions each. Following [2], the images are split into 29,000, 1000 and 1000 for training, validation and testing. As for MSCOCO dataset, there are a total of 123,287 images and every image has 5 description captions. As did in [2] [14] [15], the splits contain 113,287 images for training, 5000 for validation and 5000 for testing. Especially for MSCOCO, the final results are obtained either by averaging over 5 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.

$$Rsum = \underbrace{R@1 + R@5 + R@10}_{\text{image retrieval}} + \underbrace{R@1 + R@5 + R@10}_{\text{text retrieval}} \tag{16}$$

B. Implementation Details

We give more detailed parameter settings and model settings for our DSRAN. For global-wise feature maps extraction, the raw image is randomly cropped and resized to 224×224. And the output feature map size n is set to $7 \times 7 = 49$. For region-wise object features extraction we simply use the features given by [26] and the num of regions k is 100. Both these two kinds of features share the same dimension D_o which is 2048. As for texts, we use a pre-trained BERT-base [9] model and the embedding dimension D_w is 768. The text encoder is finetuned while parameters of visual encoders ResNet152 and Faster-RCNN are frozen. The embedded latent space dimension D_e is set to 1024. Repeating times x is 2 and 1 for MSCOCO and Flickr30K respectively.

Experiments are performed on at least two NVIDIA 1080Ti GPUs with the batch size setting to 320 for MSCOCO and 128 for Flickr30K. We train the model with an Adam optimizer [27] with a warmup rate of 0.1 for 20 epochs. The learning rate is set to $2e$-5 at first and decline by 10 times every 10 epochs.

C. Comparative experiments with state-of-the-art methods

We compare our DSRAN model with current state-of-the-art methods. They are divided into two kinds, i)global-wise visual representations methods VSE++ [2], MTFN [30] and TOD-Net [28], ii)region-wise visual representations methods SCAN [14], PFAN [29] and VSRN [15]. Further, our method is denoted as global-region unified visual representations. It should be noticed that TOD-Net uses the 24-layer BERT-large model rather than our 12-layer BERT-base model. Results from a single model or two-model ensemble are both recorded here. When conducting the ensemble scheme, the similarity scores from two already trained models are averaged.

- Results on MSCOCO

As shown in Table I, the highest performance of each metric is made bold. Our DSRAN outperforms other methods whether using an ensemble or not. For the 1K test set, our model exceeds the current best TOD-Net [28] with a BERT-large text encoder against our BERT-base encoder by 0.9% and 1.5% 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, we outnumber the state-of-the-art VSRN [15] by 9.0 considering the $Rsum$ metric. The above outperforming proves the effectiveness of our dual semantic relationship attention scheme focusing on the unified global-region visual representations learning.

- 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. Compared to the previous best model VSRN [15], we increase 7.0% on text retrieval and 9.7% on image retrieval ($R@1$), with a great improvement on Rsum metric (23.3). It is noticed that region-wise methods like SCAN [14] or VSRN perform better than global-wise counterparts VSE++ [2] or MTFN [30] which means the success of learning object-level semantic relationship for image-text matching. However, introducing unified global-region visual relationship learning further boosts the performances, which is our main contribution.

TABLE III
PERFORMANCE GAIN FROM SRR AND JRR. WE RUN THIS ABLATION STUDY ON FLICKR30K DATASET.

Modules		**Image-to-Text**			**Text-to-Image**			
SRR	**JRR**	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
		70.0	91.9	96.4	52.7	81.6	89.5	482.1
✓		72.1	92.4	96.8	55.8	83.5	89.3	489.9
	✓	72.7	92.6	96.6	56.1	83.7	89.9	491.6
✓	✓	**74.6**	**93.8**	**97.2**	**57.8**	**84.8**	**90.6**	**498.7**

V. ABLATION STUDY AND ANALYSIS

In this section, firstly we do several ablation studies considering the dual semantic relationship enhancement schemes used in our model, i)separate semantic relationship module, ii)joint semantic relationship module.

A. Effectiveness of Both Semantic Relationship Modules

There are two main semantic relationship modules in our DSRAN, the separate semantic relationship module (referred to as SSR) and the joint semantic relationship module (referred to as JSR). We perform four ablation experiments on Flickr30K [17] test set with or without the modules. The baseline configuration is to remove both two modules and merely fuse global and regional features with the fusion precoess. As shown in Table III, models with only SRR or JRR outperform the baseline. The best performance is found on the last line indicating that these two modules interact well with each other. More specifically, the performance gain comes from nowhere but the dual semantic relationship attention schemes, the first of which is the SRR contributing to the object-level relationship learning. JRR plays an important role in dealing with the unification of global and regional features thus learning the global-regional semantic relationship.

B. Analysis on Graph Attention Module

In both the two semantic relationship modules, we apply GATs [35] to enhance whether the object-wise relationship or the object-global wise relationship. By constructing two separate fully-connected graphs for global features and regional features respectively, the model successfully learns the relationship-enhanced features. Here the graph attention module can repeat for x times for deeper attention. We perform experiments to find out the best repeating times for the two datasets. As seen in Fig. 3, we adopt $x = 1\&2$ for Flickr30K and MSCOCO respectively.

Then the use of unified graph attention in the joint semantic relationship module helps to construct a graph which contains both regional features and global features. The attention progress of both features boost the relationship between objects and global concepts thus making visual representations better interact with words representations.

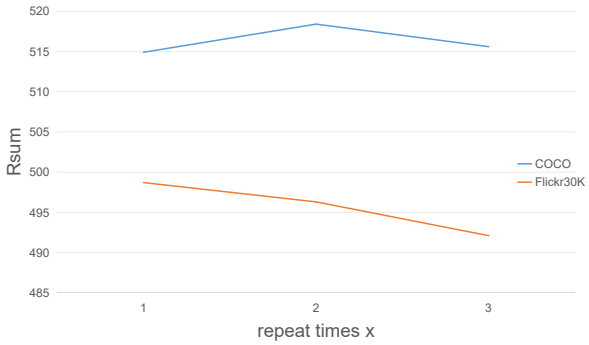

Fig. 3. How Rsum goes as x grows for two datasets.

VI. Conclusion

In this paper, we focus on the visual semantic relationship learning for enhanced image-text matching. Further a dual semantic relationship attention network (DSRAN) with different kinds of attention mechanisms applied to capture both the object-level semantic relationship and global-regional semantic relationship. The learned dual-relationships-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 the successful target-oriented designs of our model and such a model outperforms previous methods on the image-text matching task on the two widely used datasets MSCOCO and Flickr30K. Further we analyze the two main modules targeting at dual semantic relationships learning. In the future, we are looking forward to introducing scene graphs to semantic relationship learning and applying this kind of dual semantic relationship learning to more cross-modal tasks.

References

[1] R. Kiros, R. Salakhutdinov and RS. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models." arXiv preprint arXiv:1411.2539 ,2014.

[2] F. Faghri, DJ. Fleet, JR. Kiros. "Vse++: Improved visual-semantic embeddings." arXiv preprint arXiv:1707.05612 2.7, 2017: 8.

[3] B. Wang, Y. Yang, X. Xu, A. Hanjalic. "Adversarial cross-modal retrieval." Proceedings of the 25th ACM international conference on Multimedia. ACM, 2017.

[4] K. Simonyan, A. Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.

[5] A. Krizhevsky, I. Sutskever, GE. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[6] K. He, X. Zhang, S. Ren, J. Sun. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[7] S. Hochreiter, J. Schmidhuber. "Long short-term memory." Neural computation 9.8, 1997: 1735-1780.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078, 2014.

[9] J. Devlin, MW. Chang, K. Lee, K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018.

[10] A. Radford, K. Narasimhan, T. Salimans. "Improving language understanding by generative pre-training." URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf, 2018.

[11] Z. Yang, Z. Dai, Y. Yang, J. Carbonell. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." arXiv preprint arXiv:1906.08237, 2019.

[12] S. Ren, K. He, R. Girshick, J. Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[13] P. Anderson, X. He, C. Buehler. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[14] KH. Lee, X. Chen, G. Hua, H. Hu. "Stacked cross attention for image-text matching." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[15] K. Li, Y. Zhang, K. Li, Y. Li, Y. Fu. "Visual semantic reasoning for image-text matching." Proceedings of the IEEE International Conference on Computer Vision. 2019.

[16] TY. Lin, M. Maire, S. Belongie, J. Hays, P. Perona. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[17] BA. Plummer, L. Wang, CM. Cervantes. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." Proceedings of the IEEE international conference on computer vision. 2015.

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International Journal of Computer Vision 123.1, 2017: 32-73.

[19] TN. Kipf, M. Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907, 2016.

[20] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena. "Self-attention generative adversarial networks." arXiv preprint arXiv:1805.08318, 2018.

[21] S. Wang, R. Wang, Z. Yao, S. Shan. "Cross-modal Scene Graph Matching for Relationship-aware Image-Text Retrieval." arXiv preprint arXiv:1910.05134, 2019.

[22] R. Zellers, M. Yatskar, S. Thomson. "Neural motifs: Scene graph parsing with global context." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[23] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei. "Generating semantically precise scene graphs from textual descriptions for improved image retrieval." Proceedings of the fourth workshop on vision and language. 2015.

[24] J. Deng, W. Dong, R. Socher, LJ. Li, K. Li. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[25] A. Vaswani, N. Shazeer, N. Parmar. "Attention is all you need." Advances in neural information processing systems. 2017.

[26] L. Zhou, H. Palangi, L. Zhang, H. Hu, JJ. Corso. "Unified vision-language pre-training for image captioning and vqa." arXiv preprint arXiv:1909.11059, 2019.

[27] DP. Kingma, J. Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.

[28] T. Matsubara. "Target-Oriented Deformation of Visual-Semantic Embedding Space." arXiv preprint arXiv:1910.06514, 2019.

[29] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li. "Position focused attention network for image-text matching." arXiv preprint arXiv:1907.09748, 2019.

[30] T. Wang, X. Xu, Y. Yang, A. Hanjalic, HT. Shen. "Matching Images and Text with Multi-modal Tensor Fusion and Re-ranking." Proceedings of the 27th ACM International Conference on Multimedia. ACM, 2019.

[31] A. Karpathy, L. Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[32] ZM. Chen, XS. Wei, P. Wang. "Multi-Label Image Recognition with Graph Convolutional Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[33] DR. Hardoon, S. Szedmak, J. Shawe-Taylor, and John. Shawe-Taylor. "Canonical correlation analysis: An overview with application to learning methods." Neural computation 16.12, 2004: 2639-2664.

[34] Y. Huang, Q. Wu, C. Song. "Learning semantic concepts and order for image and sentence matching." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[35] P. Veličković, G. Cucurull, A. Casanov. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.

[36] Z. Li, F. Ling, C. Zhang and H. Ma, "Combining Global and Local Similarity for Cross-Media Retrieval," in IEEE Access, vol. 8, pp. 21847-21856, 2020.