Multi-Level Visual-Semantic Alignments with Relation-Wise Dual Attention Network for Image and Text Matching

Zhibin Hu[1*], **Yongsheng Luo**[1*], **Jiong Lin**[1*], **Yan Yan**[2] and **Jian Chen**[1†]
[1]School of Software Engineering, South China University of Technology, China
[2]Department of Computer Science, The University of Iowa, USA
{huzhibinscut, lysluoyongsheng, linjiong.tt, yanyan.tju}@gmail.com, ellachen@scut.edu.cn

Abstract

Image-text matching is central to visual-semantic cross-modal retrieval and has been attracting extensive attention recently. Previous studies have been devoted to finding the latent correspondence between image regions and words, e.g., connecting key words to specific regions of salient objects. However, existing methods are usually committed to handle concrete objects, rather than abstract ones, e.g., a description of some action, which in fact are also ubiquitous in description texts of real-world. The main challenge in dealing with abstract objects is that there is no explicit connections between them, unlike their concrete counterparts. One therefore has to alternatively find the implicit and intrinsic connections between them. In this paper, we propose a relation-wise dual attention network (RDAN) for image-text matching. Specifically, we maintain an over-complete set that contains pairs of regions and words. Then built upon this set, we encode the local correlations and the global dependencies between regions and words by training a visual-semantic network. Then a dual pathway attention network is presented to infer the visual-semantic alignments and image-text similarity. Extensive experiments validate the efficacy of our method, by achieving the state-of-the-art performance on several public benchmark datasets.

1 Introduction

Image and text matching is central to visual-semantic cross-modal retrieval (e.g., given a sentence query to find matched images for visual description and given an image query to retrieve related sentences for semantic description). The pivotal challenge of such tasks is to explore a strategy that can well infer the visual-semantic alignments for measuring the image-text similarity. However, due to the existing huge visual-semantic discrepancy of cross-modal data, it is challenging to infer the accurate visual-semantic alignments.

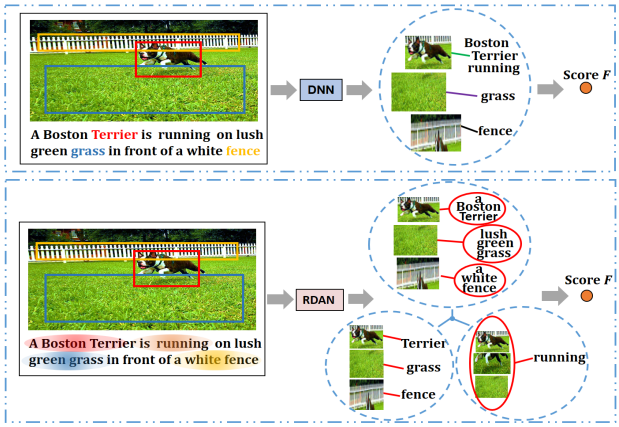

Figure 1: Illustration about the difference between current visual-semantic interaction DNN based methods and our RDAN model, which shows that our method can explore not only salient objects and key words, but also the latent relations between abstract objects.

Recently, with the dramatic development of deep learning, various deep learning based matching methods have been proposed and achieved promising performance. Many methods adopt a visual-semantic embedding based strategy [Frome *et al.*, 2013; Gong *et al.*, 2014], which maps both images and texts into a common representation space. Thus the image-text similarity can be directly measured through this common space. However, using the whole images and texts information, visual-semantic embedding based strategy ignores the importance of local visual-semantic similarities of fine-grained image-text pairs (regions and words) and blends some redundant information (useless regions).

For further solving the visual-semantic discrepancy, recently studies [Karpathy and Fei-Fei, 2015] utilize a visual-semantic interaction based strategy to measure image-text similarity. Visual-semantic interaction based strategy captures local visual-semantic similarities by comparing fine-grained image-text pairs (regions and words), and then aggregates these local similarities to obtain global image-text similarity. Moreover, considering the contribution of each local similarity is different in constructing global image-text similarity, attention mechanism [Xu *et al.*, 2015] has been introduced to discriminate the importance of each local similarity and infer the visual-semantic alignments [Huang *et al.*, 2017].

[*]Authors contributed equally
[†]Corresponding author

Such interaction-based attention methods can infer more accurate visual-semantic alignments and achieve promising performance.

However, existing methods are usually limited to explore the alignments between image regions containing concrete salience and key words, and ignoring the abstract objects, such as action. Specifically, when matching an image and a text, one not only focuses on the correspondence between concrete objects and key words, but also considers the latent relations between abstract objects. As shown in Figure 1, verbs (e.g., 'running') with discriminative information in text are often ignored, since it may be unclear which region in the image it can be matched with. People usually extract the latent relations between abstract objects when matching an image and a text. Therefore, it may motivate multi-level visual-semantic alignments for image-text matching. In reality, images and texts are often more complicated than the example of Figure 1, which makes it difficult to accurately measure visual-semantic similarity by only using the fine-grained image-text pairs. Thus, how to capture the latent visual-semantic relations and infer accurate multi-level visual-semantic alignments are the keys to further breaking the boundaries between vision and language.

To address the issues mentioned above, we introduce a novel relation-wise dual attention network (RDAN) that can infer multi-level visual-semantic alignments for measuring image-text similarity. Specifically, we first maintain an overcomplete set of image-text pairs and calculate the local similarity of all the fine-grained pairs. Based on these local similarity, we then use a visual-semantic relation CNN model to extract the latent relation by capturing the local correlations and long-term dependencies between regions and words. Furthermore, we infer the visual-semantic alignments and calculate the image-text similarity through the learned information. Concretely, we propose a dual pathway attention network, which uses a row-wise attention operation and a column-wise attention operation to obtain the attended text-level features and image-level features for measuring the image-text similarity. To summarize, the main contributions of our work are as follows:

- We propose a novel relation-wise dual attention network to explore not only the local fine-grained similarities, but also the latent visual-semantic relations, which can provide rich complementary information for inferring visual-semantic alignments and measuring image-text similarity.
- Detailed visualization of the attention results validates that our model effectively infers the accurate visual-semantic alignments.
- Experimental results conducted on two publicly available datasets demonstrate the effectiveness of the proposed model.

2 Related Work

We first review the visual-semantic embedding based methods and visual-semantic interaction based methods. Then we discuss recent advance in combining semantic-enhanced strategies into image and text matching methods.

Over the past several years, many matching methods have been proposed to explore an accurate common embedding representation. [Frome *et al.*, 2013] proposed the first deep visual-semantic embedding method. [Kiros *et al.*, 2014] combined CNN [Krizhevsky *et al.*, 2012] and LSTM [Hochreiter and Schmidhuber, 1997] to learn a common representation space. [Wang *et al.*, 2016] utilized cross-view and within-view constraints to learn structure-preserving representations. In addition, deep canonical correlation analysis [Klein *et al.*, 2015; Yan and Mikolajczyk, 2015] is used as the objective function for representation learning. Under the similar objective, [Lev *et al.*, 2016] used Fisher Vector to learn more discriminative representations. Recently, [Niu *et al.*, 2017] presented a model that maps phrases, regions, sentences and images into a shared embedding space. [Faghri *et al.*, 2017] introduced hard negatives into triplet loss function to improve the embedding learning. [Gu *et al.*, 2018] proposed incorporating generative objectives for cross-view feature embedding learning.

Visual-semantic interaction based methods utilize the local similarities of fine-grained image-text pairs to aggregate the global similarity. [Karpathy and Fei-Fei, 2015] proposed the first visual-semantic interaction based framework. [Plummer *et al.*, 2015] considered the region-to-phrase correspondences for learning the similarity. Since each fine-grained pair plays a different role in calculating the global similarity score, attention mechanism [Xu *et al.*, 2015] is applied to image-text matching problem. [Nam *et al.*, 2017] proposed a dual attentional network to capture the fine-grained interplay between regions and words. [Huang *et al.*, 2017] presented a context-modulated attention scheme to selectively attend to a pair of instances appearing in the image and sentence. [Lee *et al.*, 2018] proposed a stacked cross attention network, which learns all the possible alignments between regions and words.

Recently, some studies explored to utilize semantic-enhanced strategies to learn the visual-semantic alignments. [Qi *et al.*, 2018] constructed pairwise combinations between regions/words to represent the correlations. Then the authors utilized KNN method to model these correlations for learning visual-semantic alignments. [Huang *et al.*, 2018] used a multi-regional multi-label CNN to extract semantic concepts, and then used images and semantic concepts to generate sentence representation for measuring image-text similarity. Different from existing approaches, our method aims to directly discover the latent relations between regions and words, and to learn more discriminative visual-semantic alignments for inferring image-text similarity.

3 Our RDAN Approach

In order to capture the latent relations and infer more discriminative visual-semantic alignments, we propose a relation-wise dual attention network (RDAN). The architecture of RDAN is shown in Figure 2. Specifically, given an input image and a related text, we first map the image and text into a set of region features and a set of word features respectively. Then we use a visual-semantic relation CNN model to capture the latent relations between regions and words. Finally, we present a dual pathway attention network to infer

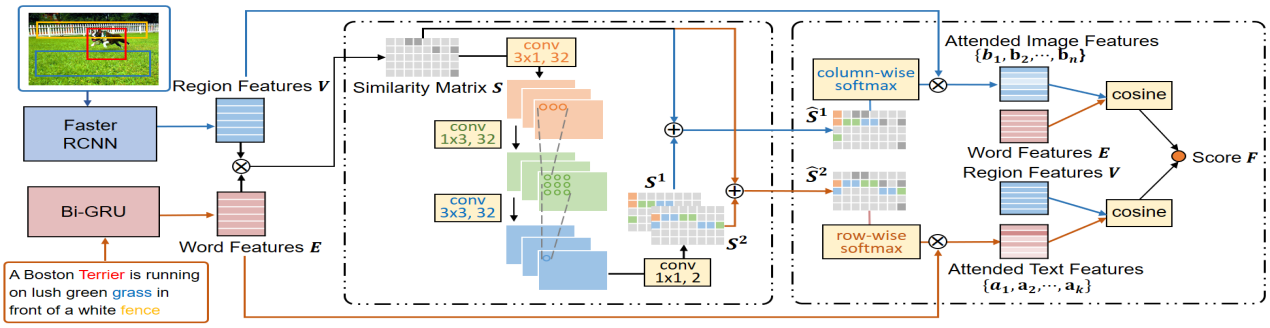

Figure 2: Architecture of the proposed relation-wise dual attention network.

the visual-semantic alignments and to calculate the image-text similarity. Next, we will introduce RDAN from the following aspects: 1) input representation of images and texts as an over-complete set, 2) details of visual-semantic relation CNN model and dual pathway attention network and 3) model learning and implementation details of RDAN.

3.1 Input Representation

In order to construct an over-complete image-text-pair set and make image and text comparable, we map both data from their own spaces to a D-dimensional common space.

For an image, we aim to represent it with a set of region features $V = (\mathbf{v}_1, \cdots, \mathbf{v}_k) \in \mathbb{R}^{D\times k}$ where k is the number of regions. Following [Anderson *et al.*, 2018], we detect the region features for each image with a Faster R-CNN [Ren *et al.*, 2015] model. We adopt the Faster R-CNN model in conjunction with a 101-layers ResNet. For each obtained region feature vector $\mathbf{r}_i \in \mathbb{R}^{d_v}$, we use a fully-connect layer to transform $\mathbf{r}_i$ into a D-dimensional vector

$$\mathbf{v}_i = W_v\mathbf{r}_i + b_v, \tag{1}$$

where W_v is a parameter matrix, $b_v \in \mathbb{R}$ is a bias vector. Thus, we can represent an image with a set of region vectors $V = \{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$. All extracted image regions are ordered by their confidence score. In our experiments, we set $D = 1024$ and $k = 36$.

For a text, we extract the same D-dimensional word features $E = (\mathbf{e}_1, ..., \mathbf{e}_n) \in \mathbb{R}^{D\times n}$ where n denotes the number of words. For i-th word in the text, we first encode it to a one-hot vector $\mathbf{o}_i$, and then map it into a 300-dimensional vector as follows

$$\mathbf{x}_i = W_o\mathbf{o}_i + b_o \tag{2}$$

Next, we use a bi-directional GRU [Bahdanau *et al.*, 2014] to learn the word representation. The bi-directional GRU contains a forward GRU, which scans the text from the first word to the last word, and a backward GRU, which scans the text by a reverse order

$$\overrightarrow{\mathbf{h}_i} = \overrightarrow{\text{GRU}}(\mathbf{x}_i);\ \ \overleftarrow{\mathbf{h}_i} = \overleftarrow{\text{GRU}}(\mathbf{x}_i),\ \ i \in \{1, ..., n\}, \tag{3}$$

We finally represent the word feature by averaging $\overrightarrow{\mathbf{h}_i}$ and $\overleftarrow{\mathbf{h}_i}$,

$$\mathbf{e}_i = \frac{\overrightarrow{\mathbf{h}_i} + \overleftarrow{\mathbf{h}_i}}{2},\ \ i \in \{1, ..., n\}. \tag{4}$$

Thus, we can obtain a set of word vectors E.

3.2 Visual-Semantic Relation CNN Model

Based on the region features V and word features E, we intend to capture the latent visual-semantic relations. Inspired by the convolution neural network, which can effectively extract the relationships between pixels and construct the expressive representation hierarchically. Therefore, we aim to use a multi-layer CNN to capture the local correlations and long-term dependencies between regions and words.

Considering the particularity of the visual-semantic interaction, we design a novel visual-semantic relation CNN model. Specifically, we first compute the cosine similarity matrix S for all region-word pairs,

$$s_{ij} = \frac{\mathbf{v}_i^\top \mathbf{e}_j}{\|\mathbf{v}_i\|\|\mathbf{e}_j\|}, i \in \{1, ..., k\}, j \in \{1, ..., n\}. \tag{5}$$

Here, s_{ij} represents the similarity between the i-th region and the j-th word. k denotes the number of regions and n denotes the number of words. We use this similarity matrix as the input of visual-semantic relation CNN model. Unlike existing methods that applying CNN to image processing [Krizhevsky *et al.*, 2012] or text processing [Pang *et al.*, 2016], each element of the input denotes a pixel or a correlation between words, each element in our model means the interaction information of each region-word pair. Then we introduce three different convolution kernels to expand the perceptual field for the regions and words. Concretely, we adopt a 3×1 convolution kernel to capture the latent relationships for uni-gram word and tri-gram regions, a 1×3 convolution kernel for uni-gram region and tri-gram words, and a 3×3 convolution kernel for tri-gram regions and tri-gram words.

The first convolutional layer filters the $k \times n$ input with 32 kernels of size 3×1. The second convolutional layer has 32 kernels of size 1×3 connected to the outputs of the first convolutional layer. The third convolutional layer has 32 kernels of size 3×3. The ReLU non-linearity is applied to the output of each convolutional layer. The operation of each layer is shown as follows,

$$S_t = \text{ReLU}(\text{Conv}(S_{t-1})). \tag{6}$$

Then we filter the output of third convolutional layer with 2 kernel of size 1×1 to obtain the matrices S^1 and S^2. Note that we remove the down-sampling operations (e.g., max-pooling) to avoid the information loss and keep the dimension of the

similarity matrix. Through the hierarchical convolution operations, we can capture the latent visual-semantic relations from local to global. Finally, we merge the learned matrices and the original similarity matrix S

$$\hat{S}^1 = S^1 \oplus S, \qquad \hat{S}^2 = S^2 \oplus S, \tag{7}$$

where $\oplus$ is an element-wise plus operation. We regard the similarity matrix S as a residual term to preserve the fine-grained interaction information. Thus, the visual-semantic relation CNN model can capture not only the local fine-grained similarities, but also the latent relations, which can provide rich complementary information for inferring visual-semantic alignments.

3.3 Dual Pathway Attention Network

Armed by the learned visual-semantic relations, we can measure more accurate image-text similarity. To this end, we propose a dual pathway attention network, which can infer the importance of all the words to each region and infer the importance of all the regions to each word. Then we construct attended image-level feature to each word and construct attended text-level feature to each region for measuring image-text similarity.

For i-th region, we fitst use a row-wise attention operation on $\hat{S}^2$ to calculate the weight of each word to i-th region. Then we extract a corresponding attended text-level vector through a weighted combination of word representations,

$$\mathbf{a}_i = \sum_{j=1}^{n} \alpha_{ij}\mathbf{e}_j, \;\; \alpha_{ij} = \frac{\exp(\lambda \hat{s}_{ij}^2)}{\sum_{j=1}^{n}\exp(\lambda \hat{s}_{ij}^2)}, \tag{8}$$

where λ is the inversed temperature of the softmax function [Chorowski *et al.*, 2015]. Similarly, we can obtain a corresponding attended image-level vector for j-th word through a column-wise attention operation on $\hat{S}^1$ as follows,

$$\mathbf{b}_j = \sum_{i=1}^{k} \beta_{ij}\mathbf{v}_i, \;\; \beta_{ij} = \frac{\exp(\lambda \hat{s}_{ij}^1)}{\sum_{i=1}^{k}\exp(\lambda \hat{s}_{ij}^1)}. \tag{9}$$

Through the above dual pathway attention operations, we can obtain an image-level vector as context for each word and obtain a text-level vector as context for each region. Then we calculate the region relevance and word relevance as follows,

$$\begin{aligned} R_r(\mathbf{v}_i, \mathbf{a}_i) &= \frac{\mathbf{v}_i^\top \mathbf{a}_i}{\|\mathbf{v}_i\|\|\mathbf{a}_i\|}, i \in \{1, ..., k\}, \\ R_w(\mathbf{e}_j, \mathbf{b}_j) &= \frac{\mathbf{e}_j^\top \mathbf{b}_j}{\|\mathbf{e}_j\|\|\mathbf{b}_j\|}, j \in \{1, ..., n\}. \end{aligned} \tag{10}$$

Here, region relevance denotes the similarity between i-th region and corresponding text-level vector, word relevance denotes the similarity between the j-th word and corresponding image-level vector. Finally, the visual-semantic similarity between image I and text T is calculated as follows:

$$F(I,T) = (1-\mu)\times\frac{\sum_{i=1}^{k} R_r(\mathbf{v}_i, \mathbf{a}_i)}{k} + \mu \times \frac{\sum_{j=1}^{n} R_w(\mathbf{e}_j, \mathbf{b}_j)}{n}, \tag{11}$$

where μ is a hyper-parameter, which controls the balance between region relevance and word relevance.

3.4 Model Learning

We utilize the hinge-based triplet loss as the objective function, which is widely used in image-text matching field. Given a positive image-text pair (I, T), the objective function is as follows,

$$\begin{aligned} L(I,T) = &\sum_{\hat{T}} \max(0, m - F(I,T) + F(I,\hat{T})) \\ &+ \sum_{\hat{I}} \max(0, m - F(I,T) + F(\hat{I},T)) \end{aligned} \tag{12}$$

where m is a tuning margin, $F(I,T)$ denotes the similarity score of matched image I and text T, $F(I,\hat{T})$ is the score of mismatched image I and text $\hat{T}$, and vice-versa with $F(\hat{I},T)$. The above function considers all negative text $\hat{T}$ and all negative image $\hat{I}$.

However, using all the negative samples will lead to expensive computation. A common approach is to select a fixed number of mismatched pairs. Following [Faghri *et al.*, 2017], we focus on the hardest negatives. For a positive image-text pair (I, T), the hardest negative is given by $\hat{I}_h = \text{argmax}_{g \neq I} F(g, T)$ and $\hat{T}_h = \text{argmax}_{d \neq T} F(I, d)$. Thus, the objective function for optimizing our model is defined as follows,

$$\begin{aligned} L_h(I,T) = &\max(0, m - F(I,T) + F(I,\hat{T}_h)) \\ &+ \max(0, m - F(I,T) + F(\hat{I}_h,T)) \end{aligned} \tag{13}$$

All modules of our proposed RDAN excepting for the image regions extraction can constitute a whole deep network, which can be trained in an end-to-end manner.

3.5 Implementation Details

Our RDAN approach is implemented by Pytorch. For images, we adopt a Faster R-CNN model to extract region features. We set the intersection over union (IOU) threshold as 0.7. We extract the top k vectors of the last pooling layer for each image, and then use a fully-connect layer to transform the region vectors into 1024-dimensional vectors. For texts, we use the bidirectional GRU to encode a text into a set of word features. The word feature can be obtained by averaging the first 1024-dimensional output of forward GRU and backward GRU. Other parameters are empirically set as follows: $k = 36$, $\mu = 0.1$, $\lambda = 4$ and $m = 0.2$.

4 Experiments

To demonstrate the effectiveness of the proposed RDAN model, we conduct extensive experiments on the visual-semantic cross-modal retrieval task. We use the Flickr30k [Plummer *et al.*, 2015] and MS-COCO [Lin *et al.*, 2014] datasets. These datasets contain 31,000 and 123,287 images respectively and each image has 5 captions. For Flickr30K, we follow [Karpathy and Fei-Fei, 2015] to use 1,000 images for validation, 1,000 images for testing and the rest for training. For MS-COCO, we follow [Lee *et al.*, 2018] to use 5,000 images for validation, 5,000 images for testing and the rest for training. We report the results by averaging over 5 folds

Method	Sentence Retrieval			Image Retrieval			Sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DVSA [Karpathy and Fei-Fei, 2015]	22.2	48.2	61.4	15.2	37.7	50.5	235.2
HM-LSTM [Niu *et al.*, 2017]	38.1	-	76.5	27.7	-	68.8	-
DSPE [Wang *et al.*, 2016]	40.3	68.9	79.9	29.7	60.1	72.1	351.0
SM-LSTM [Huang *et al.*, 2017]	42.5	71.9	81.5	30.2	60.4	72.3	358.8
CRAN [Qi *et al.*, 2018]	38.1	70.8	82.8	38.1	71.1	82.6	383.5
2WayNet [Eisenschtat and Wolf, 2017]	49.8	67.5	-	36.0	55.6	-	-
DAN [Nam *et al.*, 2017]	55.0	81.8	89.0	39.4	69.2	79.1	413.5
VSE++ [Faghri *et al.*, 2017]	52.9	-	87.2	39.6	-	79.5	-
DPC [Zheng *et al.*, 2017]	55.6	81.9	89.5	39.1	69.2	80.9	416.2
SCO [Huang *et al.*, 2018]	55.5	82.0	89.3	41.1	70.5	80.1	418.5
SCAN [Lee *et al.*, 2018]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
RDAN	**68.1**	**91.0**	**95.9**	**54.1**	**80.9**	**87.2**	**477.2**

Table 1: Comparison results of sentence retrieval and image retrieval on the Flickr30k dataset.

Method	Sentence Retrieval			Image Retrieval			Sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DVSA [Karpathy and Fei-Fei, 2015]	38.4	69.9	80.5	27.4	60.2	74.8	351.2
HM-LSTM [Niu *et al.*, 2017]	43.9	-	87.8	36.1	-	86.7	-
DSPE [Wang *et al.*, 2016]	50.1	79.7	89.2	39.6	75.2	86.9	420.7
SM-LSTM [Huang *et al.*, 2017]	53.2	83.1	91.5	40.7	75.8	87.4	431.7
2WayNet [Eisenschtat and Wolf, 2017]	55.8	75.2	-	39.7	63.3	-	-
VSE++ [Faghri *et al.*, 2017]	64.6	-	95.7	52.0	-	92.0	-
DPC [Zheng *et al.*, 2017]	65.6	89.8	95.5	47.1	79.9	90.0	467.9
GXN [Gu *et al.*, 2018]	68.5	-	97.9	56.6	-	94.5	-
SCO [Huang *et al.*, 2018]	69.9	92.9	97.5	56.7	87.5	**94.8**	499.3
SCAN [Lee *et al.*, 2018]	72.7	94.8	98.4	58.8	88.4	**94.8**	507.9
RDAN	**74.6**	**96.2**	**98.7**	**61.6**	**89.2**	94.7	**515.0**

Table 2: Comparison results of sentence retrieval and image retrieval on the MS-COCO dataset.

of 1K test images. We use Recall@K as the metric for evaluation, which means the correct image (sentence) is ranked within the Top-K retrieved results to the sentence query (image query). We calculate another criterion 'Sum' to evaluate the overall performance for both sentence retrieval and image retrieval.

4.1 Comparison with State-of-the-art Methods

We compare our model with several state-of-the-art models on Flickr30k and MS-COCO datasets in Table 1 and Table 2. We can observe that our proposed model outperforms all baselines on both datasets. Especially for R@1, which is crucial to measure accuracy, our model achieves the best performance. For example, our best results at R@1 are 74.6 and 61.6 for sentence retrieval and image retrieval on MS-COCO dataset, which improves 2.6% on sentence retrieval and 4.7% on image retrieval comparing to current state-of-the-art. Similar observations can be obtained from other metrics (*e.g.*, R@5 and R@10). In addition, we notice that our models have significant improvements in the overall performance. The results demonstrate that our RDAN model can effectively infer visual-semantic alignments and accurately measure the image-text similarity.

4.2 Ablation Studies

We aim to validate the contribution of each component of our model by carrying out some ablation experiments. Specifically, we intend to answer the following questions: 1) Is the dual pathway attention structure effective? 2) Whether the visual-semantic relation CNN model is helpful to learn visual-semantic alignments? 3) Is the residual term (similarity matrix S) useful? 4) Does the hard negatives technique is effective?

We first train RDAN with $\mu = 0$ or $\mu = 1$, which means that RDAN only uses row-wise/column-wise attention operation to infer similarity score. We also train RDAN with one pathway attention structure ($\text{RDAN}_{\text{one-path}}$). This model generates one learned matrix, which is used to infer similarity score by using column-wise and row-wise attention operations. We then train RDAN without the visual-semantic relation CNN model ($\text{RDAN}_{\text{no-cnn}}$), which denotes RDAN only uses the original similarity matrix. Next, we train RDAN without the residual term S in the visual-semantic relation CNN model ($\text{RDAN}_{\text{no-res}}$). Finally, we train RDAN without using the hard negatives technique ($\text{RDAN}_{\text{no-hard}}$).

The experimental results are shown in Table 3. We observe that the dual pathway attention structure is effective to learn more accurate visual-semantic alignments for inferring similarity score. The performance will degrade dramatically when only using row-wise/column-wise attention operation. In addition, one pathway attention structure is also difficult to achieve satisfactory results. Besides, we can see the residual term S is useful, since it contains the interaction information of fine-grained image-text pairs, which is important for inferring visual-semantic alignments. Further more, the hard negatives technique can significantly improve the performance.

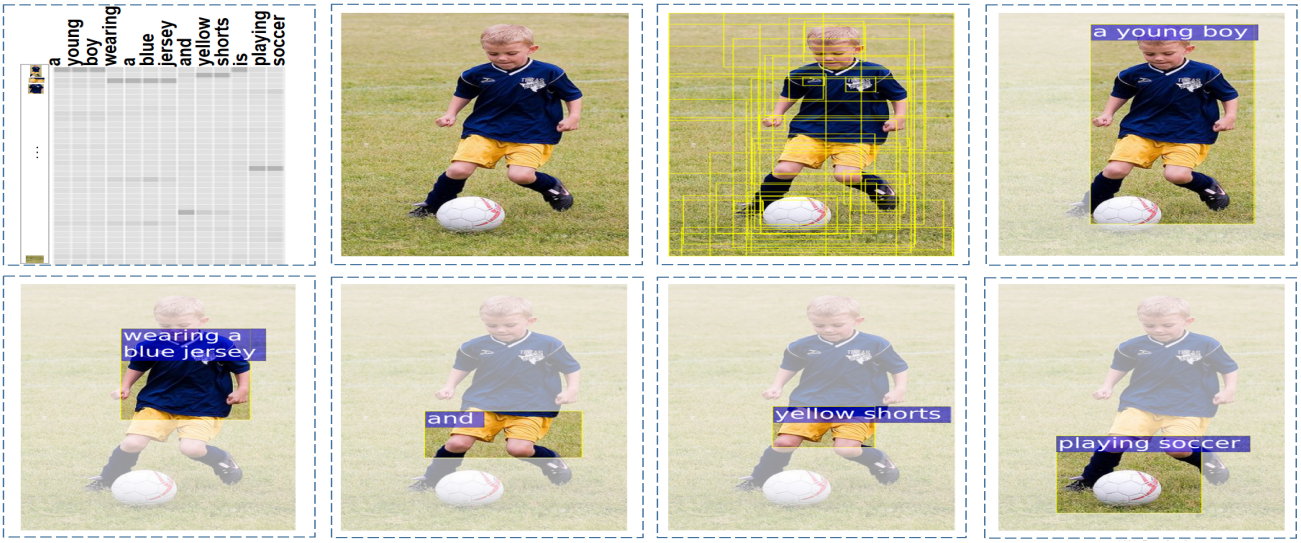

Figure 3: Visualization of the image regions with respect to each word in the sentence description.

Method	Flickr30K dataset							MS-COCO dataset						
	Sentence Retrieval			Image Retrieval			Sum	Sentence Retrieval			Image Retrieval			Sum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
RDAN$_{\mu=0}$	63.5	90.4	94.8	51.7	79.0	86.7	466.1	71.0	95.6	98.4	58.2	87.0	92.3	502.5
RDAN$_{\mu=1}$	62.6	87.6	94.6	46.6	76.0	82.7	450.1	62.7	91.9	96.1	51.4	84.5	92.1	478.7
RDAN$_{\text{one-path}}$	62.8	89.6	95.4	49.7	78.1	85.8	461.4	57.8	90.6	97.0	55.4	87.4	93.9	482.1
RDAN$_{\text{no-cnn}}$	65.9	89.1	94.6	44.5	74.7	84.8	453.6	68.4	94.5	98.1	50.5	84.9	92.7	489.1
RDAN$_{\text{no-res}}$	64.5	90.0	95.0	49.3	78.4	85.3	462.5	59.4	91.6	97.1	56.1	87.0	93.9	485.1
RDAN$_{\text{no-hard}}$	63.7	88.2	94.6	48.7	76.9	84.3	456.4	68.9	94.1	98.0	55.8	86.6	93.3	496.7
RDAN	**68.1**	**91.0**	**95.9**	**54.1**	**80.9**	**87.2**	**477.2**	**74.6**	**96.2**	**98.7**	**61.6**	**89.2**	**94.7**	**515.0**

Table 3: Ablation experiment result on Flick30K and MS-COCO dataset

4.3 Analysis of Visualizing Attention Results

In order to show the discriminative ability and interpretability of our learned visual-semantic alignments, we visualize the learned most important region corresponding to each word. Specifically, we first calculate attention weights among each word and all regions on the learned information matrix. Then we visualize the most important region with respect to each word based on the weights.

In Figure 3, we first show the learned information matrix for the selected image with a sentence 'A young boy wearing a blue jersey and yellow shorts is playing soccer'. We then display the original image and the image with extracted region bounding box. We match the most important region for each word in the rest sub-figures. We can observe that the words 'a', 'young' and 'boy' are mapped to the same image region and the words 'wearing', 'a', 'blue' and 'jersey' are mapped to another same image region. These observations prove that our model can effectively explore not only concrete objects and key words, but also the latent relations between abstract objects. Such multi-level visual-semantic alignments are more in line with human behavior when matching images and texts. In this way, we can learn the more expressive image-level/text-level representation for each word/region, which is useful for measuring image-text similarity.

5 Conclusion

This paper aims to deal with abstract objects in image-text matching. Unlike the concrete objects, abstract ones lack the explicit connection between text and image region, requiring alternative ways to explore the intrinsic connection. Therefore, we propose a relation-wise dual attention network to capture the latent relations and infer visual-semantic alignments. We first use a visual-semantic relation network to learn latent correlations over a over-complete set of image-text pairs. We then present a dual pathway attention network to obtain more expressive region/word representations and to measure image-text similarity. In addition, we provide visualization analysis to show how RDAN can give more discriminative and interpretability to such vison-language models. We perform experiments on cross-modal retrieval tasks and the results demonstrate the effectiveness of the proposed model by achieving significant performance improvements.

Acknowledgments

The authors thank the anonymous reviewers for their helpful comments. This work was supported by the Guangdong special branch plans young talent with scientific and technological innovation (2016TQ03X445) and the Guangzhou science and technology planning project (201904010197).

References

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Chorowski *et al.*, 2015] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *NeurIPS*, pages 577–585, 2015.

[Eisenschtat and Wolf, 2017] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, pages 4601–4611, 2017.

[Faghri *et al.*, 2017] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129, 2013.

[Gong *et al.*, 2014] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014.

[Gu *et al.*, 2018] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, pages 7181–7189, 2018.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Huang *et al.*, 2017] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, pages 2310–2318, 2017.

[Huang *et al.*, 2018] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, pages 6163–6171, 2018.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[Klein *et al.*, 2015] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, pages 4437–4446, 2015.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[Lev *et al.*, 2016] Guy Lev, Gil Sadeh, Benjamin Klein, and Lior Wolf. Rnn fisher vectors for action recognition and image annotation. In *ECCV*, pages 833–850. Springer, 2016.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Nam *et al.*, 2017] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, pages 299–307, 2017.

[Niu *et al.*, 2017] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *ICCV*, pages 1881–1889, 2017.

[Pang *et al.*, 2016] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. In *AAAI*, 2016.

[Plummer *et al.*, 2015] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015.

[Qi *et al.*, 2018] Jinwei Qi, Yuxin Peng, and Yuxin Yuan. Cross-media multi-level alignment with relation attention network. In *IJCAI*, 2018.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[Wang *et al.*, 2016] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013, 2016.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[Yan and Mikolajczyk, 2015] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, pages 3441–3450, 2015.

[Zheng *et al.*, 2017] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535*, 2017.