Joint Global and Co-Attentive Representation Learning for Image-Sentence Retrieval

Shuhui Wang[1], Yangyu Chen[1,2], Junbao Zhuo[1,2], Qingming Huang[1,2,*], Qi Tian[3,4]
[1] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China.
[2] University of Chinese Academy of Sciences, Beijing, 100049, China.
[3] Huawei Noah's Ark Lab.
[4] University of Texas at San Antonio, Texas, USA.
wangshuhui@ict.ac.cn,{yangyu.chen,junbao.zhuo}@vipl.ict.ac.cn,qmhuang@ucas.ac.cn,tian.qi1@huawei.com

ABSTRACT

In image-sentence retrieval task, correlated images and sentences involve different levels of semantic relevance. However, existing multi-modal representation learning paradigms fail to capture the meaningful component relation on word and phrase level, while the attention-based methods still suffer from component-level mismatching and huge computation burden. We propose a Joint Global and Co-Attentive Representation learning method (JGCAR) for image-sentence retrieval. We formulate a global representation learning task which utilizes both intra-modal and inter-modal relative similarity to optimize the semantic consistency of the visual/textual component representations. We further develop a co-attention learning procedure to fully exploit different levels of visual-linguistic relations. We design a novel softmax-like bi-directional ranking loss to learn the co-attentive representation for image-sentence similarity computation. It is capable of discovering the correlative components and rectifying inappropriate component-level correlation to produce more accurate sentence-level ranking results. By joint global and co-attentive representation learning, the latter benefits from the former by producing more semantically consistent component representation, and the former also benefits from the latter by back-propagating the contextual information. Image-sentence retrieval is performed as a two-step process in the testing stage, inheriting advantages on both effectiveness and efficiency. Experiments show that JGCAR outperforms existing methods on MSCOCO and Flickr30K image-sentence retrieval tasks.

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Computing methodologies** → Learning linear models; Learning latent representations;

KEYWORDS

Image-sentence retrieval, co-attentive representation, cross-modal representation learning, joint learning.

ACM Reference Format:
Shuhui Wang, Yangyu Chen, Junbao Zhuo, Qingming Huang, Qi Tian. 2018. Joint Global and Co-Attentive Representation Learning for Image-Sentence Retrieval. In MM '18: 2018 ACM Multimedia Conference, Oct. 22–26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240535

1 INTRODUCTION

Vision and language are two important aspects in understanding the world. Research endeavor is motivated by breaking the boundaries between the two in image-sentence matching [13, 20, 22, 37, 41], image captioning [14, 21, 22] and visual question answering (VQA) [19, 40, 43]. The key to bridging vision-language gap is to find a good metric that accurately measures the semantic image-sentence similarity, and based on which the semantically similar images and sentences can be properly associated. Owning to recent achievement in deep learning [6, 28, 35], image-sentence retrieval (a.k.a., cross-modal retrieval [37]) is built on top of modality-specific representation learning modules, thus promising performance has been reported on benchmark evaluation [20, 38].

As a straight-forward way of image-sentence retrieval, a global joint embedding space is learned which maximizes document-level (*i.e.*, sentence-level or image-level) semantic correlation using various linear or nonlinear mapping functions [1, 8, 11, 13, 22, 29, 37, 38]. Accordingly, documents from different modalities are represented with low dimensional vectors, and their distances/similarities that reflect their semantic relations are intermediately measured for relevance ranking. To deal with the evident information discrepancy and modality heterogeneity between visual and textual contents, some modality specific mapping functions are employed in global embedding learning, *e.g.*, CNN for image or LSTM for sentence. The learned representations suffer from much irrelevant information brought by useless visual regions and words in the whole document. For example, some meaningless background visual regions or meaningless words may be mistakenly used to calculate the global representation, and they may dominate the correlation contribution.

In fact, the correlated images and sentences involve different levels of component semantic relevance. For example, words describe objects in image, phrases describe attributes or activities

[*] Corresponding author.

MM '18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240535

of objects, and the whole sentence expresses the topic of the entire image. Beyond document-level correlation, word-level mapping [5, 32, 33], phrase-level correspondence [14, 23, 26, 48] and their combination [20] are utilized towards a more comprehensive multi-level correlation aggregation. Recently, attention [19, 40, 43] was introduced into vision-language task to identify “where to look" (*i.e.*, spatial maps) and “which words to listen to" (*i.e.*, meaningful words/phrases) in image-text pairs. A bottom-up question-image co-attention model [19] is proposed to identify the hierarchical correlative components. Compared to global representation learning paradigms, these approaches are able to explore meaningful component correspondence that contributes significantly to the cross-modal relation.

However, it is usually assumed that semantically correlated components in image and sentence depend on each other for attention-based models, which means that different image-sentence pairs may have different component-wise correspondence. In this case, there may not exist a unique meaningful (*i.e.*, attentive) component set for a single document without knowing its cross-modal counterpart [19, 22]. Searching for true component-wise correspondence/co-attention is very time-consuming, leading to inefficiency in cross-modal correlation learning. Besides, due to the diverse component-level content representation, incorrect matching between words/phrases and visual regions may be inappropriately introduced during the co-attention learning process, which results in inaccurate bottom-up image-sentence correlation modeling.

To address both effectiveness and efficiency issues, we investigate image-sentence retrieval by a collaboration of global representation learning and co-attention learning. We formulate a global representation learning task which utilizes both intra-modal and inter-modal relative similarity to optimize the semantic consistency of the global component-based visual and textual representations. The global representation not only enhances the overall component-wise semantic matching accuracy, but it can also be used as the first-stage searching to identify candidate component pairs that are highly possible to be true correspondence. To further capture the inter-modal relation, we propose a co-attention learning procedure which fully exploits different levels of image-sentence matching relations. The upper level correlation depends on lower level one, *e.g.*, phrase is generated by selected meaningful words and its corresponding visual region by smaller meaningful visual regions correlated with the selected words. We design a novel softmax-like bi-directional cross-modal ranking loss, and by minimizing the loss the co-attentive representation for image-sentence similarity computation is learned. It is capable of discovering the correlative components and their contextual relation in image and sentence, and producing a more accurate sentence-level ranking results by rectifying inappropriate component-level correlation.

With our Joint Global and Co-Attentive Representation learning method (JGCAR), the subtle and implicit semantic relations between images and sentences are fully exploited. The latter task benefits from the former task by producing more semantically consistent component representation, and the former also benefits from the latter by back-propagating the co-attention to components. Image-sentence retrieval is performed as a two-step process in testing stage, where a set of cross-modal candidates are first identified by comparing similarity on global representation, and the final ranking results are obtained by ranking on co-attentive representation similarities. The number of trials for time consuming co-attentive image-sentence correlation inference is significantly reduced, thus both effectiveness and efficiency are gained. Promising results have been achieved by our approach on MSCOCO and Flickr30K image-sentence retrieval tasks, which demonstrates the remarkable accuracy of JGCAR in identifying sentence-level and component-level correlation.

2 THE PROPOSED METHOD

We are given a training set $O = \{\mathbf{V}, \mathbf{S}\}$. Let $\mathbf{V} = \{V_1, \ldots, V_{N_v}\}$ denote the visual features of N_v images. $\mathbf{S}^r = \{S_1^r, \ldots, S_{N_t}^r\}$ denotes the textual features of N_t sentences, where $r = w$ denotes the word-level feature, $r = p$ denotes the phrase-level feature and $r = s$ denotes the sentence-level feature. Without loss of generality, we use $i(i') \in \{1, \ldots, N_v\}$ to denote the index of image and $j(j') \in \{1, \ldots, N_t\}$ to denote the index of sentence. We present the joint learning framework in Figure 1, which performs global and co-attentive representation jointly for image-sentence retrieval. Note that our method involves a joint learning in the training stage, and a two-stage retrieval in the testing stage.

2.1 Visual/Textual Component Representation

Textual features. Given a sentence j with T words, $G_j = [g_j(1), \ldots, g_j(T)]$ denotes its 1-hot encoding representation, where $g_j(t)$ is the feature vector of the t-th word, we obtain sentence representation from low-level word features progressively. We first embed words into a low-dimensional space through an embedding matrix to obtain its word-level feature $S_j^w = [s_j^w(1), \ldots, s_j^w(T)]$ as

$$s_j^w(t) = \mathbf{W}_e g_j(t), \ t \in \{1, 2, \cdots, T\} \tag{1}$$

where $\mathbf{W}_e$ is weight parameters that can be optimized towards specific task. Similar as [19], at each word location t, we use three CNN-like convolution filters to compute the phrase-level features, which have the sizes of unigram, bigram and trigram, to calculate inner product of word vectors. The t-th convolutional output using window size c is computed by

$$\hat{s}_{j,c}^p(t) = \tanh(\mathbf{W}_c s_j^w(t : t + c - 1) + b_c), \ c \in \{1, 2, 3\} \tag{2}$$

where $\mathbf{W}_c$ and b_c are weight parameters. The word-level features S_j^w are 0-padded before feeding into bigram and trigram convolutions, by which we ensure the same lengths of different feature sequences after convolution. After obtaining the convolution outputs, we obtain phrase-level features $S_j^p = [s_j^p(1), \ldots, s_j^p(T)]$ by max-pooling operation across feature vectors at each word location t as

$$s_j^p(t) = \max(\hat{s}_{j,1}^p(t), \hat{s}_{j,2}^p(t), \hat{s}_{j,3}^p(t)), \ t \in \{1, 2, \ldots, T\} \tag{3}$$

To fully exploit the semantic information of the sequential features $s_j^p(t)$, we feed the phrase-level vectors $s_j^p(t)$ of words into bi-directional LSTM [7]. At location t, the d-dimensional sentence-level feature $s_j^s(t)$ is calculated by adding hidden vectors from the forward and backward LSTMs

$$s_j^s(t) = \text{Bi-LSTM}(s_j^p(t), \phi_m), \ t \in \{1, 2, \ldots, T\} \tag{4}$$

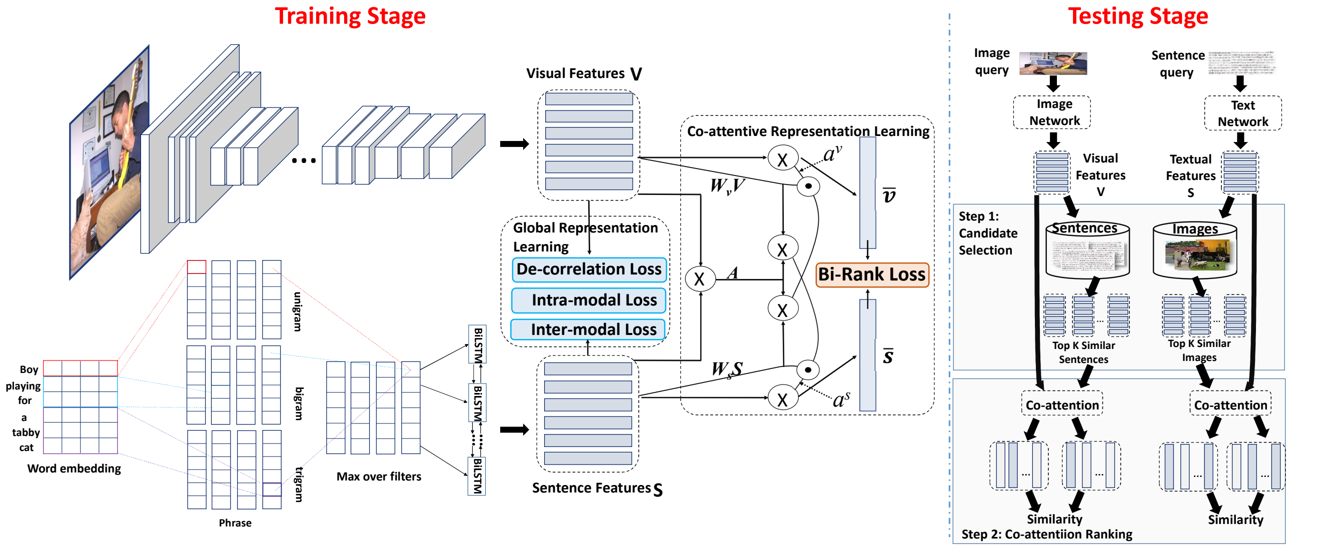

Figure 1: The framework of our approach. In the training stage, the global representation learning task optimizes the visual representation V and sentence representation S by minimizing inter-modal loss, intra-modal loss and de-correlation loss; the co-attentive representation learning task optimizes V and S and learns the element-wise weights jointly by minimizing the bi-rank loss to produce the co-attentive representation $\bar{v}$ and $\bar{s}$ for each image-sentence pair. In the testing stage, given a specific query type, the top most similar cross-modal candidates are first returned by comparing the similarity on V and S, then the final ranking score is produced on the co-attentive representation between the query and each of the cross-modal candidates.

where ϕ_m denotes the parameters of bi-directional LSTM. Then we obtain the feature for a given sentence j as $S_j^s = [s_j^s(1), \ldots, s_j^s(T)]$. We denote $S_j^s = f_t(G_j, \theta_s)$, where θ_s includes $\mathbf{W}_e$, $\mathbf{W}_c$, b_c and ϕ_m.

Visual features. We apply VGG-16 [28] for extracting visual feature from an original image. Each image i is first re-scaled to be 448×448 pixels. Then we select the feature from the last pooling layer to preserve the spatial information of the original image, so the feature has a dimension of $512 \times 14 \times 14$. We denote the feature as $R_i = [r_i(1), ..., r_i(N)]$, where $N = 14 \times 14$ is the number of regions in the raw image and 512 is the dimension of the feature vector $r_i(n)$ of each region. A single layer perceptron is used to embed each 512-dim feature vector into a new vector that has the same dimension d as the sentence vector.

$$v_i(n) = \tanh(\mathbf{W}_I r_i(n) + b_I) \quad (5)$$

where $\mathbf{W}_I \in \mathcal{R}^{d \times 512}$ is an embedding matrix and $b_I \in \mathcal{R}^d$ is a bias term. For simplicity, the feature of image i is denoted by $V_i = [v_i(1), \ldots, v_i(N)] = f_v(I_i, \theta_v)$ where θ_v includes the parameters of CNN, $\mathbf{W}_I$ and b_I.

2.2 Global Representation Learning

Given the visual representation $V_i \in \mathcal{R}^{d \times N}$ produced by f_v for an image i and textual representation $S_j^s \in \mathcal{R}^{d \times T}$ produced by f_t for a sentence j, we aim to learn the global representations for each image and sentence. To this end, we employ the intra-modal and inter-modal semantic relation on the image-sentence dataset.

Specifically, we use $\widehat{V}_i$ and $\widehat{S}_j^s$ to denote the d dimensional features derived by *Global Average Pooling* on V_i and S_j^s, respectively. Denote $\Omega_{ij} = \widehat{V}_i^\top \widehat{S}_j^s$, we apply bi-directional triplet loss [38] to model the inter-modal relation as

$$\begin{aligned} \mathcal{J}_1 = & \sum_{i,j,j^-} \max\left[0, m - \Omega_{ij} + \Omega_{ij^-}\right] \\ & + \sum_{j,i,i^-} \max\left[0, m - \Omega_{ij} + \Omega_{i^-j}\right] \end{aligned} \quad (6)$$

where m denotes the margin in triplet loss. (i, j) is correlated image-sentence pair, i^- denotes the uncorrelated image to j and j^- the uncorrelated sentence to i. It is easy to find that minimizing this loss will drag correlated image-sentence pairs to be closer, and push uncorrelated image-sentence pairs away. Therefore, we can preserve the relative semantic similarities of cross-modal instances.

Also, good representation should also have good discriminative abilities in their own modality to preserve semantic information. The semantic relation in each modality are beneficial to improve the performance of cross-modal retrieval. By enforcing the global representations for semantically similar instance to be similar to each other, the global representations are also endowed with more discriminative power. To this end, we add the intra-modal pairwise embedding loss for image modality and textual modality, respectively. For image modality, based on the visual representation $V_i, i = 1, \ldots N_v$, the constraint is formulated as

$$\mathcal{J}_2 = \sum_{i,i^+,i^-} \max\left[0, m - \Omega_{ii^+} + \Omega_{ii^-}\right] \quad (7)$$

where $\Omega_{ii^+} = \widehat{V}_i^\top \widehat{V}_{i^+}$. Similarly, for textual modality, based on the textual representation of sentences $S_j^s, j = 1, \ldots N_t$, the constraint is formulated as

$$\mathcal{J}_3 = \sum_{j,j^+,j^-} \max\left[0, m - \Omega_{jj^+} + \Omega_{jj^-}\right] \quad (8)$$

where $\Omega_{jj^+} = (\widehat{S}_j^s)^\top \widehat{S}_{j^+}^s$. In global representation learning, if some dimensions have high correlation, there will be redundant information in between. Therefore, to reduce the correlations among feature dimensions, we design a de-correlation constraint on both modalities. Given a batch of $N_v^b(\ll N_v)$ training images and $N_t^b(\ll N_t)$ training sentences, the constraint is formulated as

$$\mathcal{J}_4 = \frac{1}{2}\left(||\mathbf{C}^v||_F^2 - ||diag(\mathbf{C}^v)||_F^2\right) + \frac{1}{2}\left(||\mathbf{C}^t||_F^2 - ||diag(\mathbf{C}^t)||_F^2\right) \tag{9}$$

where $C^v(d_1, d_2) = \frac{1}{N_v^b}\sum_{k=1}^{N_v^b}(V_k(d_1,:) - \mu_{d_1})^\top(V_k(d_2,:) - \mu_{d_2})$, $d_1, d_2 = 1, \ldots, d$. $\mathbf{C}^v$ is the co-variance of different dimensions on V over the image data batch, and $\mu_{d_1} = \frac{1}{N_v^b}\sum_{k=1}^{N_v^b} V_k(d_1,:)$. Similarly, $\mathbf{C}^t$ is the co-variance of different dimensions on S^s over the sentence batch. The de-correlation enforces that different dimensions should be de-correlated to each other as much as possible on each batch. Therefore, the redundancy between dimensions can be suppressed, leading to more representative feature representation. Note that our de-correlation constraint is designed on *matrix co-variance* rather than vector co-variance as in [42]. The key difference between the two is that the spatial location information is preserved in the matrix co-variance. The covariance is only measured between features in the same spatial location in images or sentences, which avoids inappropriate loss on spatial context during the feature de-correlation procedure.

The overall global representation learning objective function is

$$\min_{\mathbf{U},\mathbf{S}^s} \mathcal{J}_\mathcal{G} = \min_{\mathbf{U},\mathbf{S}^s}(\mathcal{J}_1 + \lambda_2\mathcal{J}_2 + \lambda_3\mathcal{J}_3) + \lambda_4\mathcal{J}_4 \tag{10}$$

where $0 \le \lambda_2, \lambda_3, \lambda_4 \le 1$. We set $\lambda_2 = \lambda_3 = \lambda_4 = 1$ which guarantees good performance.

2.3 Co-attentive Representation Learning

Co-attention. Based on $V_i \in \mathcal{R}^{d\times N}$ and $S_j^s \in \mathcal{R}^{d\times T}$ for image and sentence, we propose to learn the co-attentive feature representation, which generates visual and textual attention maps simultaneously. We first define a common embedding for both image and sentence features. Then we link the embedded image and sentence features through calculating the correlation matrix, which is the weighted combined dot product between image and sentence features at all pairs of image locations and sentence locations [40]. The affinity matrix $A \in \mathcal{R}^{T\times N}$ is calculated by,

$$A_{ji} = \left(S_j^s\right)^\top \mathbf{W}_b V_i = \left(f_t(G_j, \theta_s)\right)^\top \mathbf{W}_b f_v(I_i, \theta_v) \tag{11}$$

where $\mathbf{W}_b \in \mathcal{R}^{d\times d}$ contains the attention embedding weights for both visual features V and textual features S. We consider this affinity matrix A_{ji} as a common feature of sentence j and image i. This joint operation concurrently guides the visual and textual attentions, which can make two attentions to closely cooperate with each other. We feed the image feature V_i and sentence feature S_j^s through a single neural network to generate the hidden states of image and sentence:

$$\begin{aligned} H_{i|j}^v &= \tanh(\mathbf{W}_v V_i + b_v) \odot \tanh(\mathbf{W}_s S_j^s A_{ji}) \\ H_{j|i}^s &= \tanh(\mathbf{W}_s S_j^s + b_s) \odot \tanh(\mathbf{W}_v V_i A_{ji}^\top) \end{aligned} \tag{12}$$

where $\mathbf{W}_v$, $\mathbf{W}_s \in \mathcal{R}^{l\times d}$, $b_v \in \mathcal{R}^{l\times N}$ and $b_s \in \mathcal{R}^{l\times T}$ are weight parameters. $\odot$ is element-wise multiplication. From Eqn. 12, we know that the affinity matrix A_{ji} can transform textual attention space to visual attention space (vice versa for $A_{ji}^\top$).

Then a softmax function is used to generate attention distributions over regions of image and words of sentence:

$$\begin{aligned} a_{i|j}^v &= \text{softmax}(w_{hv}^\top H_{i|j}^v + b_{hv}) \\ a_{j|i}^s &= \text{softmax}(w_{hs}^\top H_{j|i}^s + b_{hs}) \end{aligned} \tag{13}$$

where $w_{hv}, w_{hs} \in \mathcal{R}^l$ are the embedding parameters. $a^v \in \mathcal{R}^N$ and $a^s \in \mathcal{R}^T$ are the attention probabilities of each image region v_n and word s_t, respectively.

Based on the obtained attention weights, the visual and textual co-attentive features are computed as weighted sum within the image features V_i and sentence features S_j^s as follows:

$$\bar{v}_{i|j} = \sum_{n=1}^{N} a_{i|j}^v(n)v^n, \; \bar{s}_{j|i} = \sum_{t=1}^{T} a_{j|i}^s(t)s^s(t) \tag{14}$$

In image-sentence matching tasks, we need to compare numerous images and sentences. To facilitate effective and efficient cross-modal similarity learning, we proposes a novel comparison method that evaluates the relative ranking scores between given query sample and a certain retrieved sample, by which more disparities between retrieved samples can be achieved in contrasts. Based on the co-attentive representations $\hat{v}$ and $\hat{s}^s$, the relative ranking score between given query sample $\hat{v}$ and retrieved sample $\hat{s}$ is measured as:

$$\alpha(\bar{v}_{i|j}|\bar{s}_{j|i}) = \text{cosine}(\bar{v}_{i|j}, \bar{s}_{j|i}) = \frac{\bar{v}_{i|j} \cdot \bar{s}_{j|i}}{\|\bar{v}_{i|j}\|\|\bar{s}_{j|i}\|} \tag{15}$$

where $\bar{v}_{i|j}$ and $\bar{s}_{j|i}$ are the learned co-attentive features of image and sentence, respectively. $\|\cdot\|$ denotes the norm operator. Due to the task-specific property of image-sentence matching, the relative ranking score with respect to the above image-sentence similarity should be considered in two directions, *i.e.*, the image-to-sentences direction and sentence-to-images direction.

Bi-rank loss. Given an input image $\bar{v}_i$, we use its corresponding sentence $\bar{s}_{j\to i}$ as a positive sample. To obtain representative non-matching pairs, we select the top M' most dissimilar sentences and top M' most dissimilar images in each mini-batch as a negative sentence set $\mathcal{S}_i^-$ and a negative image set $\mathcal{V}_i^-$, respectively. This overall loss can effectively explore both cross-modal and intra-modal relations. Similar as [17], a softmax-like formula is used to compute the ranking score for both cross-modal comparison set $\mathcal{S}_i$ and intra-modal comparison set $\mathcal{V}_i$:

$$\begin{aligned} &P(\bar{v}_i|\bar{s}, \bar{v}, \mathcal{S}_i, \mathcal{V}_i^-) = \\ &2 - \frac{\exp(\gamma\alpha(\bar{v}_i|\bar{s}_{j\to i}))}{\sum_{\bar{s}\in\mathcal{S}_i}\exp(\gamma\alpha(\bar{v}_i|\bar{s}))} - \frac{\exp(\gamma\alpha(\bar{v}_i|\bar{s}_{j\to i}))}{\sum_{\bar{v}\in\mathcal{V}_i}\exp(\gamma\alpha(\bar{v}_i|\bar{v}))} \end{aligned} \tag{16}$$

where $\mathcal{S}_i = \bar{s}_{j\to i} \cup \{\mathcal{S}_i^-\}$ denotes the set of input sentences to be ranked, and $\mathcal{V}_i = \bar{v}_i \cup \{\mathcal{V}_i^-\}$. γ is set empirically during experiments. Similar to the Boltzmann exploration [34] in reinforcement learning, small γ enforces all retrieved sentences to be nearly e-quiprobable, while large γ increases the probability score of the positively related sentence and decreases the scores of negatively related sentence. The search scope of true matching sample will be shrunken with a larger value of γ such that the probability of

Table 1: Bidirectional retrieval results on MSCOCO.

Task / Method	Image-to-Sentences			Sentence-to-Images		
	R@1	R@5	R@10	R@1	R@5	R@10
Mean vector [15]	33.2	61.8	75.1	24.2	56.4	72.4
CCA$_{FH}$ [15]	37.7	66.6	79.1	24.9	58.8	76.5
CCA$_{FGH}$ [15]	39.4	67.9	80.9	25.1	59.8	76.6
DVSA [14]	38.4	69.9	80.5	27.4	60.2	74.8
m-RNN-VGG [21]	41.0	73.0	83.5	29.0	42.2	77.0
mCNN [20]	42.8	73.1	84.1	32.6	68.6	82.8
SPE [38]	50.1	79.7	89.2	39.6	**75.2**	**86.9**
GRL$_{d=256}$	45.6	78.5	86.9	37.8	70.7	80.6
SGCAR$_{d=256}$	51.5	80.6	89.7	39.2	71.2	82.5
JGCAR$^{v1}_{d=256}$	49.2	78.6	87.1	38.0	70.6	80.1
JGCAR$^{v2}_{d=8}$	28.1	60.2	73.7	24.6	56.7	71.3
JGCAR$^{v2}_{d=16}$	32.4	65.2	78.5	28.9	61.4	76.2
JGCAR$^{v2}_{d=256}$	52.4	82.3	90.4	39.5	74.6	85.3
JGCAR$_{d=8}$	31.8	63.7	75.6	26.5	59.3	74.9
JGCAR$_{d=16}$	36.3	67.8	80.1	31.6	63.7	78.6
JGCAR$_{d=256}$	**52.7**	**82.6**	**90.5**	**40.2**	74.8	85.7

Table 2: Bidirectional retrieval results on Flickr30K.

Task / Method	Image-to-Sentences			Sentence-to-Images		
	R@1	R@5	R@10	R@1	R@5	R@10
DCCA [41]	27.9	56.9	68.2	26.8	52.9	66.9
mCNN [20]	33.6	64.1	74.9	26.2	56.3	69.6
m-RNN-VGG [21]	35.4	63.8	73.7	22.8	50.7	63.1
SDT-RNN [29]	9.6	29.8	41.1	8.9	29.8	41.1
GHF [16]	35.0	62.0	73.8	25.0	52.7	66.0
HF [23]	36.5	62.2	73.3	24.7	53.4	66.8
SPE [38]	40.3	68.9	79.9	29.7	60.1	72.1
DAN(VGG) [22]	41.4	73.5	82.5	31.8	61.7	**72.5**
GRL$_{d=256}$	34.4	64.6	76.3	29.7	52.6	64.7
SGCAR$_{d=256}$	42.3	73.6	81.0	33.4	57.5	65.9
JGCAR$^{v1}_{d=256}$	43.2	71.4	76.8	38.0	56.4	65.3
JGCAR$^{v2}_{d=8}$	27.0	48.1	53.8	24.1	47.9	55.6
JGCAR$^{v2}_{d=16}$	32.0	54.6	63.5	26.0	50.6	59.7
JGCAR$^{v2}_{d=256}$	44.6	74.8	81.2	35.1	61.5	71.8
JGCAR$_{d=8}$	30.3	50.6	54.4	25.9	49.8	57.7
JGCAR$_{d=16}$	34.7	56.9	65.8	28.4	53.1	61.5
JGCAR$_{d=256}$	**44.9**	**75.3**	**82.7**	**35.2**	**62.0**	72.4

correct matching is increased. The sentence-to-images rank loss can be defined in similar way where M and M' are set to be identical to the image-to-sentences rank loss. Our bi-rank loss $\mathcal{J}_C$ is the sum of the two directional rank losses. Unlike previous works [22, 38], we directly use negative intra-modal and inter-modal pairs without searching for extra positively intra-modal pairs.

2.4 Joint Learning

The overall joint learning objective function is

$$\mathcal{J} = \beta \mathcal{J}_{\mathcal{G}} + (1 - \beta)\mathcal{J}_C \tag{17}$$

where $0 \leq \beta \leq 1$ denotes the weight of each task, and we heuristically set $\beta = 0.5$ under all circumstances. It can be easily optimized by stochastic gradient descent optimizer family. By analyzing Eqn. 10 we can see that, $\mathcal{J}_{\mathcal{G}}$ can be optimized with respect to V and S^s. $\mathcal{J}_C$ can be optimized with respect to α^v, α^s, U and S^s, while α^v and α^s are also parameterized by V and S^s according to the chain rule of derivatives, as seen in Eqn. 13 and 12. Therefore, the model can be trained end-to-end. V and S^s can be seen as the *information bottleneck* of the deep representation learning architecture, which aggregate both pair-wise relative semantic relation and the contextual co-attentive information. Thus enhancing efficacy on one of the two complementary tasks benefit the other.

Also, it is reasonable to use V and S^s to conduct a first-step quick candidate search without much compromise on accuracy. The truly matched objects in other modalities given a query can be easily captured using a large number of candidate set, *e.g.*, the top 100 ranked documents. In the second step, the co-attentive similarity between the query and candidates are computed to produce a more accurate re-ranking results. Based on the two-step retrieval, the potential ability of the co-attentive representation learning can be fully exploited, since it can discover the true component correlation by filtering meaningless background components and document-level correlation.

Table 3: Bidirectional retrieval results with higher K.

Task / Method	Image-to-Sentences			Sentence-to-Images		
	R@20	R@50	R@100	R@20	R@50	R@100
	On MSCOCO					
GRL	89.3	97.4	98.3	87.0	94.6	96.8
SGCAR	92.9	98.0	98.3	92.3	96.2	96.8
JGCAR	94.2	98.0	99.1	93.8	97.0	97.8
	On Flickr30K					
GRL	78.4	85.0	90.4	69.2	82.7	90.0
SGCAR	83.2	87.1	90.4	75.7	85.3	90.0
JGCAR	83.9	88.0	90.5	76.7	86.3	91.1

3 EXPERIMENTS

3.1 Experimental setting

Datasets. Flickr30K [44] consists of 31,783 images, each of which is associated with five descriptive sentences. We follow the public splits [14, 15, 22, 23, 38]: 29,783 training, 1,000 validation and 1,000 for testing. The larger MSCOCO [18] dataset consists of 123,000 images, also associated with five sentences each. On this dataset, to be consistent with [14, 15, 38], we also report results on 1,000 test images and their corresponding sentences. We use the category information with 80 classes of MSCOCO for the global representation learning task. On Flickr30K, we extract object labels from the bounding box annotation for the intra-modal triplet loss in global representation learning. We treat visual/textual instances with no identical label to be the negatively correlated examples.

Implementation details. We develop our architecture in PyTorch framework [3]. We adopt the SGD for training GRL and Rmsprop for training co-attention, where the learning rate, momentum and weight decay are set as 4e-4, 0.99 and 1e-8, respectively. We train our model by at most 128 epochs with a mini-batch size of 200, and the training will be early stopped if the validation performance has not been improved in the last 5 epochs. All word embedding and hidden layers are 512-dimensional vectors. The

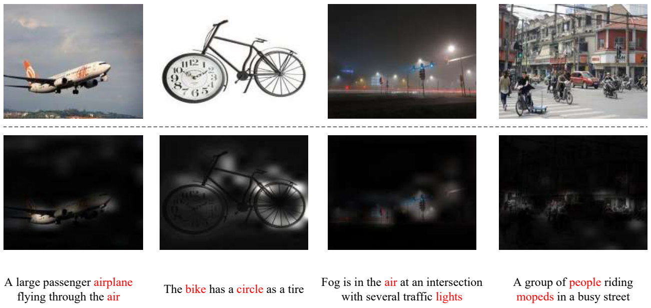
A large passenger airplane flying through the air
The bike has a circle as a tire
Fog is in the air at an intersection with several traffic lights
A group of people riding mopeds in a busy street

Figure 2: Qualitative results of image-to-sentences retrieval with attention visualization. The four query images are shown in the top row. The middle row shows the attention maps. The bottom row shows the retrieved most correlated sentences.

dimension parameter d is set to be 256 on both MSCOCO and Flickr30K to ensure good effectiveness and efficiency tradeoff. To avoid over-fitting, we apply dropout with probability 0.5 on each layer. We set the maximum word numbers of each sentence as 20 and 15 on MSCOCO and Flickr30K, respectively. We tune the parameters of bi-ranking loss until the best performance is obtained. Consequently, the ranking loss uses the parameter setting as $\gamma = 10$ and $M' = 30$. In the testing stage, we return the top 100 most similar examples as the cross-modal candidates for consequent co-attentive similarity ranking.

Competitors. We compare JGCAR with recent popular models on image-sentence retrieval. Some models explore the semantic matching in the word-level like Mean vector [15], CCA with FV + HGLMM (CCA$_{FH}$) [15], CCA with FV + GMM + HGLMM (CCA$_{FGH}$) [15], DCCA [41], GMM + HGLMM + FV (GHF) [16] and HGLMM + FV (HF) [23], and some others in sentence level like DVSA [14], m-RNN-VGG [21], SPE [38] and SDT-RNN [29]. Specifically, mCNN (ensemble) [20] exploits the semantic association at multi-level multi-modal matching. DAN (VGG) [22] is the first study on learning multi-modal attentions for image-text matching. For better illustrating the effectiveness of JGCAR, we construct another two baseline for comparison. The first is only with global representation learning (GRL), which corresponds to $\beta = 1$ for JGCAR. The second is a separate learning of global representation and co-attentive representation (SGCAR), where two cross-modal deep networks should be constructed, one for global representation learning and the other for co-attentive learning. This baseline has the same data sampling process as the full JGCAR model, where the top 100 most similar examples are chosen for co-attentive similarity computation. We do not set JGCAR where $\beta = 0$ as the third baseline since the computation burden for image-sentence retrieval using only co-attentive similarity ranking is prohibitive. To clearly identify the influence of different loss functions in our method, we also compare with several simplified version of JGCAR. Specifically, we use JGCAR^{v1} to denote our method without the intra-modal losses $\mathcal{J}_2$ and $\mathcal{J}_3$, and use JGCAR^{v2} to denote our method without de-correlation loss $\mathcal{J}_4$. On JGCAR^{v2}, we also report the results when $d = \{8, 16\}$ in addition to the optimal setting of d, to show the influence of de-correlation on different d.

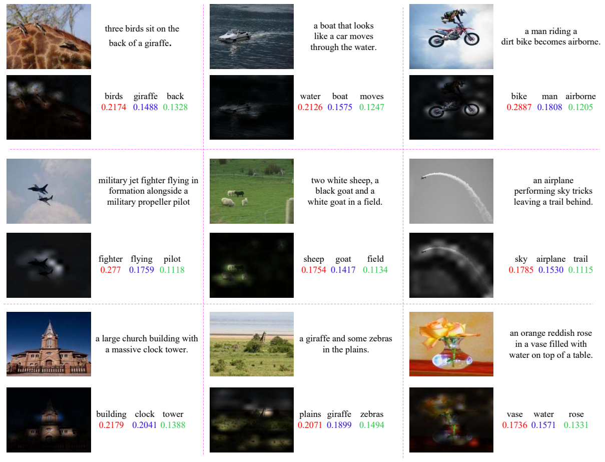
three birds sit on the back of a giraffe.
birds giraffe back
0.2174 0.1488 0.1328
a boat that looks like a car moves through the water.
water boat moves
0.2126 0.1575 0.1247
a man riding a dirt bike becomes airborne.
bike man airborne
0.2887 0.1808 0.1205
military jet fighter flying in formation alongside a military propeller pilot
fighter flying pilot
0.277 0.1759 0.1118
two white sheep, a black goat and a white goat in a field.
sheep goat field
0.1754 0.1417 0.1134
an airplane performing sky tricks leaving a trail behind.
sky airplane trail
0.1785 0.1530 0.1115
a large church building with a massive clock tower.
building clock tower
0.2179 0.2041 0.1388
a giraffe and some zebras in the plains.
plains giraffe zebras
0.2071 0.1899 0.1494
an orange reddish rose in a vase filled with water on top of a table.
vase water rose
0.1736 0.1571 0.1331

Figure 3: Visualization of some good image-sentences attention maps of JGCAR. For each instance, the original image and sentence are shown in the top row. The bottom row shows image-sentence attention maps. The top-3 salient words are shown in red, blue and green.

3.2 Image-sentence retrieval

Follow the same protocols as in [14, 15, 22, 23, 38], we report the retrieval performance of Recall@K (K=1, 5, 10) which calculates the fraction of times the ground truth match is found among the top K matches in Table 1 and 2. We highlight the best performances of each task in **boldface**.

First, in all cases, JGCAR outperforms all competitors significantly in recall@1. For example, on MSCOCO dataset, the proposed model achieves 2.6% and 0.6% higher in terms of absolute performance gain than the best compared results achieved by SPE on image-to-sentences and sentence-to-images, respectively. On Flickr dataset, JGCAR also achieves 3.5% and 3.4% higher in absolute performance than the best compared results (*i.e.*, DAN) on image-to-sentences and sentence-to-images, respectively. JGCAR outperforms others except SPE in recall@5 and recall@10 in Sentence-to-Images on MSCOCO data. The reason may be attributed to that SPE uses bounding box annotation which is more powerful than the instance-level annotation used in our approach. SPE also involves negative mining process, thus it can potentially discover more false negatives from a large number of negative candidates. On Flickr 30K data, JGCAR outperforms all others in recall@5 for both tasks, but underperforms DAN in recall@10 on sentence-to-images. The reason may be explained by the advantage of alternatively interdependent co-attention learning used in DAN. According to analysis in [19], co-attention derived alternatively may achieve a better result than a parallel co-attention mechanism as ours. Despite of that, by average the performance on both directions, JGCAR still achieves state-of-the-art performance on bidirectional image-sentence retrieval.

Second, from the performance comparison between GRL, SGCAR and JGCAR we see that our full model consistently outperforms the other two baseline approaches. The global representation learning (GRL) achieves good recall rates at all positions, and a two-step procedure even with separately trained models can further improve the performance. Remarkable performance gains are observed on

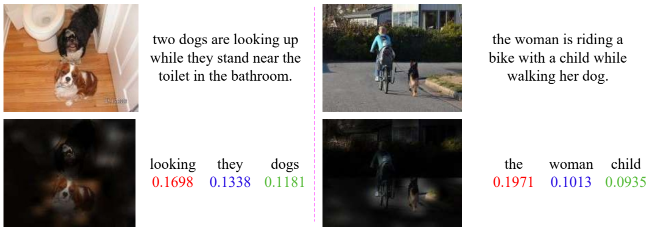
two dogs are looking up while they stand near the toilet in the bathroom.
looking 0.1698
they 0.1338
dogs 0.1181
the woman is riding a bike with a child while walking her dog.
the 0.1971
woman 0.1013
child 0.0935

Figure 4: Visualization of some poor image-sentences attention maps of JGCAR. For each instance, the original image and sentence are shown in the top row. The bottom row shows image-sentence attention maps.

JGCAR over SGCAR, which further validates that our joint learning mechanism encourages a collaborative learning on the component-wise representation V and S^s. The global representation learning improves the overall component-wise representation ability, and co-attentive representation learning endows the component-based representation with more contextual information and derives their aggregated representation effectively. More importantly, co-attentive representation is very crucial in improving the top 1 recall rate, which indicates that the co-attention representation learning is specially effective to find the best match.

Third, from the performance of different versions of JGCAR in Table 1 and 2 we can see that intra-modal losses have minor influence on R@1 but stronger impact on R@5 an R@10. $\mathcal{J}_2$ and $\mathcal{J}_3$ play complementarily with Bi-rank loss $\mathcal{J}_C$ in performance improvement. De-correlation guarantees better performance when d is small, but tends to be marginal when d is larger. Similar to orthogonal constraint, de-correlation enforces feature dimensions to be as much de-correlated as possible. It is indispensable to guarantee good performance when d is small (8 or 16) by ensuring that each feature dimension contains useful information by squeezing out as much redundancy as possible. Besides, it improves the convergence property of model training according to experimental facts.

In Table 3 we show how the recall rate changes with larger K on both datasets. The recall will increase and go close to 100% with a higher K. The R@100 of GRL and SGCAR are identical since the number of similar examples chosen for the second-stage co-attentive ranking is also 100. So R@100 of GRL and SGCAR merely depend on the performance of the first-stage ranking, and the co-attentive learning has no influence on the final performance for SGCAR. But JGCAR slightly outperforms GRL and SGCAL when $K = 100$ since the co-attention information is also injected into V and S^s, which shows the positive influence on improving the semantic consistency of V and S^s used for the first-stage ranking. The result further verifies the effectiveness of JGCAR.

3.3 Illustrative examples

We show some attention visualization results of image-to-sentences retrieval in Figure 2. We observe that JGCAR effectively detects the important semantic components appeared in both modalities. This property mainly depends on correctly identified relevant semantic description of each modality, and then minimizing the proposed ranking loss further enhances their semantic correlation. Figure 3 provides some good examples of image-sentence co-attention by our model. For all instances, it can discover the meaningful components for both image and sentence, and further describe the semantic relation of the explored correlative components. In each instance, the top-3 salient words clearly describe their related visual objects.

We give some poor results of co-attention generated by our model. For the two instances, the visual and textual attentions are somewhat inaccurate. This is mainly due to the gap between human-level cognition and the attention representation. For example, the objects of an image may have some positional relationship, *e.g.*, "riding a bike" implies that the "woman" should be located above of the "bike". Unfortunately, this relationship can hardly be learned without knowledge guidance. Moreover, objects in an image may also have affiliation relation. For example, "toilet in the bathroom" indicates that "bathroom" naturally contains "toilet". Therefore, it is our next-step work to explore these complex context by incorporating cognitive knowledge towards a more comprehensive correlation modeling for image-sentence retrieval.

3.4 Parameter sensitivity analysis

We investigate two important parameters in the proposed method, *i.e.*, the dimension number d and the scope parameter γ. Without loss of generality, parameter sensitivity analysis is conducted on training sets on both MSCOCO and Flickr30K to test how these parameters impact the performance on the validation data. For each parameter, we conduct empirical analysis by varying its value and fixing the other parameters, and then we show the performance on validation data. Figure 5 shows the Recall@K (K=1, 5, 10) scores of image-sentence matching with different parameter settings. First, the left two columns show the performance with different d on the two retrieval directions. We can see that the performance is enhanced when d becomes larger. The performance keeps the same level as $d = 256$ when d is larger than 256. Therefore, we set $d = 256$ to ensure both efficiency and effectiveness. Second, the right two columns show the performance with different scope parameter γ on the two retrieval directions. Small γ will make all retrieved samples to be nearly equiprobable, thus it is hard for each sample to find its counterpart, resulting in poor performances. With a larger γ, the performances are gradually improved. However, the performance is not improved with constantly increasing γ because this will reduce the diversity of samples. In our experiment, the gradient will become "nan" when γ is greater than 70 on both datasets. Therefore, we set $\gamma = 20$ to guarantee the robustness of cross-modal Bi-ranking.

4 RELATED WORK

The key problem in image-sentence retrieval is to measure the semantic similarity between visual and textual domains. A common idea is to learn a joint embedding space where heterogeneous features are directly comparable [24, 25, 27]. It has been proved that both the inter-modal relation and intra-modal similarity are equally important in constructing a good metric for measuring the image-text similarity [47]. Numerous nonlinear [9] or non-parametric [30] cross-modal representation learning paradigms have been proposed

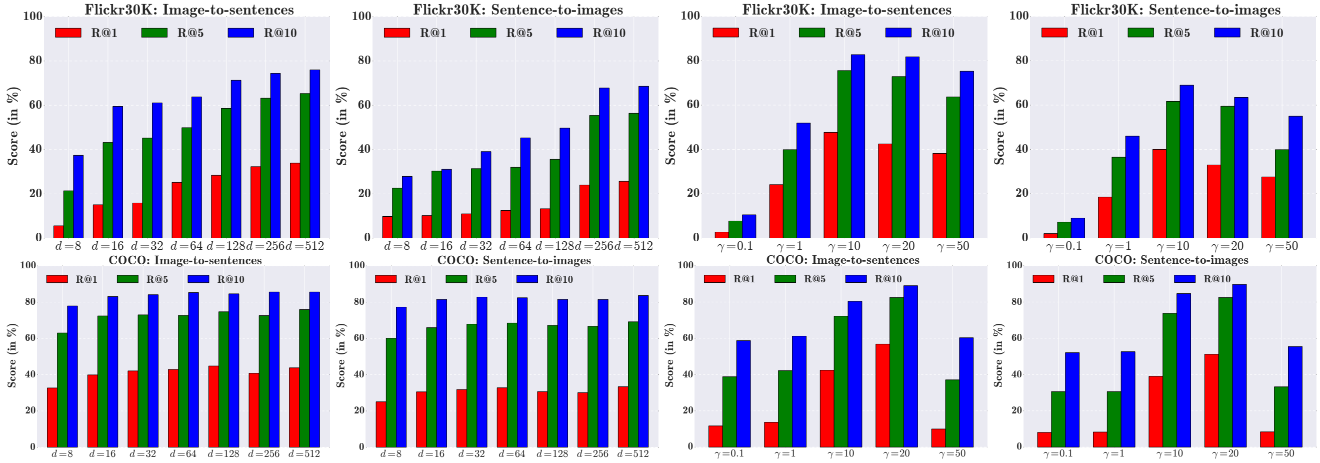

Figure 5: Parameter sensitivity analysis of dimensions d and γ on Flickr30K and MSCOCO datasets.

using both intra-modal and inter-modal relations for model learning. Cross-modal hashing, as a close relative of subspace learning, aims to transform images and texts into a joint Hamming embedding space with graph-based [12, 31], rank-based [4], alternating learning [10], online learning [39] or distributed (local) learning mechanisms [45, 46], and promising performance has been achieved on real-world multimedia search or cross-modal retrieval.

However, the domain properties and component relations have not been fully explored by previous subspace-based approaches. It may lead to limited learning capability on real world problems. To address this concern, some works focus on modeling the relation between image and annotated words [1, 5, 11, 24, 27, 32, 33, 37], and some others pay attention to image and phrases [26, 48]. These works have weakness in capturing complicated matching relation between image and sentence. Recent works have been proposed to explore the semantic relation at sentence level [8, 13, 20, 22, 29, 38]. For image-sentence retrieval, Socher *et al.* [29] adopt semantic dependency-tree recursive neural network (SDT-RNN) to embed images and sentences into the joint semantic space, where their similarity can be measured. Wang *et al.* [38] design a two-branch neural network with cross-view ranking and within-view structure preservation constraints for modeling semantic image-sentence relations. However, they neglect the local fragment of sentence and their corresponding visual patches. In contrast, Karpathy *et al.* [13] focus on a finer level matching by constructing semantic relation between sentence fragments and visual regions. Nam *et al.* [22] first utilize attention model to automatically highlight the shared semantic relations between images and sentences. Although the local inter-modal semantic relations between image regions and sentences fragments are highlighted, the global matching relations are ignored. Ma *et al.* [20] proposed a multimodal convolutional neural network (m-CNN) to construct different semantic fragments from words, and then the fragments are interacted with image at different levels. It takes different matching levels as separate steps, which ignores the intrinsic relation from words to sentence. In image captioning and visual question answering tasks, neural image caption [36], multimodal recurrent neural network (m-RNN) [21] and deep visual-semantic alignments(DVSA) [14], produce possibilities for generating caption for a given image. Deeper LSTM Question [2], dual attention networks [22] and hierarchical co-attention [19], conducts multi-modal reasoning to predict answer for a question relating to a given image.

5 CONCLUSION

We propose a joint global and co-attentive representation learning method for image-sentence retrieval. We formulate a global representation learning task which utilizes both intra-modal and inter-modal pair-wise relative relations to optimize the semantic consistency of the visual and textual component representations. We further propose a co-attention learning procedure to fully exploit different levels of visual-linguistic relations by minimizing the softmax-like bi-directional ranking loss for image-sentence similarity computation. By joint global and co-attentive representation learning, the latter benefits from the former by producing more semantically consistent component representation, and the former also benefits from the latter by back-propagating the contextual component information. Experiments show that JGCAR outperforms existing methods on MSCOCO and Flickr30K image-sentence retrieval tasks quantitatively and qualitatively. In future work, we will study how to improve the efficiency of the first stage search using global representations, and further enhance the robustness and efficacy of co-attention learning.

6 ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China: 61672497, 61332016, 61620106009, 61650202, U1636214, 61429201, in part by National Basic Research Program of China (973 Program): 2015CB351802, in part by Key Research Program of Frontier Sciences of CAS: QYZDJ-SSW-SYS013, and in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. 2013. Deep Canonical Correlation Analysis. In *ICML*. 1247–1255.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Zitnick, and D. Parikh. 2015. VQA: Visual Question Answering. In *ICCV*. 2425–2433.

[3] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. 2011. Torch7: A Matlab-like Environment for Machine Learning. In *BigLearn, NIPS Workshop*.

[4] Kun Ding, Bin Fan, Chunlei Huo, Shiming Xiang, and Chunhong Pan. 2017. Cross-Modal Hashing via Rank-Order Preserving. *IEEE TMM* 19, 3 (2017), 571–585.

[5] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. 2013. Devise: A deep Visual-Semantic Embedding Model. In *NIPS*. 2121–2129.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[7] Sepp Hochreiter, J urgen Schmidhuber, and Corso Elvezia. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[8] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR* 47 (2013), 853–899.

[9] Yan Hua, Shuhui Wang, Siyuan Liu, Anni Cai, and Qingming Huang. 2016. Cross-Modal Correlation Learning by Adaptive Hierarchical Semantic Aggregation. *IEEE TMM* 18, 6 (2016), 1201–1216.

[10] Go Irie, Hiroyuki Arai, and Yukinobu Taniguchi. 2015. Alternating Co-Quantization for Cross-Modal Hashing. In *ICCV*. 1886–1894.

[11] W. Jason, B. Samy, and U. Nicolas. 2010. Large Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings. *Machine Learning* 81 (2010), 21–35.

[12] Qing Yuan Jiang and Wu Jun Li. 2017. Deep Cross-Modal Hashing. In *CVPR*. 3270–3278.

[13] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *NIPS*. 1889–1897.

[14] Andrej Karpathy and Fei-Fei Li. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*. 3128–3137.

[15] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher Vectors Derived from Hybrid Gaussian-Laplacian Mixture Models for Image Annotation. *arXiv preprint arXiv:1411.7399* (2014).

[16] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating Neural Word Embeddings with Deep Image Representations Using Fisher Vectors. In *CVPR*. 4437–4446.

[17] Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial Ranking for Language Generation. In *NIPS*. 3155–3165.

[18] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*. 740–755.

[19] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-attention for Visual Question Answering. In *NIPS*. 289–297.

[20] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal Convolutional Neural Networks for Matching Image and Sentence. In *ICCV*. 2623–2631.

[21] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep Captioning with Multimodal Recurrent Neural Networks (M-RNN). *arXiv preprint arXiv:1412.6632* (2014).

[22] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *CVPR*. 299–307.

[23] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-phrase Correspondences for Richer Image-to-sentence Models. In *ICCV*. 2641–2649.

[24] V. Ranjan, N. Rasiwasia, and C. Jawahar. 2015. Multi-Label Cross-modal Retrieval. In *ICCV*. 4094–4102.

[25] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. 2010. A New Approach to Cross-modal Multimedia Retrieval. In *ACMMM*. 251–260.

[26] Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition Using Visual Phrases. In *CVPR*. 1745–1752.

[27] A. Sharma, A. Kumar, D. Hal, and D. Jacobs. 2012. Generalized Multiview Analysis: A Discriminative Latent Space. In *CVPR*. 2160–2167.

[28] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[29] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *TACL* 2 (2014), 207–218.

[30] Guoli Song, Shuhui Wang, Qingming Huang, and Qi Tian. 2017. Multimodal Similarity Gaussian Process Latent Variable Model. *IEEE TIP* 26, 9 (2017), 4168–4181.

[31] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. 2013. Inter-media Hashing for Large-scale Retrieval from Heterogeneous Data Sources. In *ACM SIGMOD*. 785–796.

[32] Nitish Srivastava and Ruslan Salakhutdinov. 2012. Learning Representations for Multimodal Data with Deep Belief Nets. In *ICML Representation Learning Workshop*, Vol. 79.

[33] N. Srivastava and R. Salakhutdinov. 2012. Multimodal Learning with Deep Boltzmann Machines. In *NIPS*. 2222–2230.

[34] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement Learning: An Introduction*. Vol. 1. MIT press Cambridge.

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *CVPR*. 1–9.

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *CVPR*. 3156–3164.

[37] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. 2013. Learning Coupled Feature Spaces for Cross-modal Matching. In *ICCV*. 2088–2095.

[38] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-preserving Image-text Embeddings. In *CVPR*. 5005–5013.

[39] Liang Xie, Jialie Shen, Lei Zhu, et al. 2016. Online Cross-Modal Hashing for Web Image Retrieval.. In *AAAI*. 294–300.

[40] Huijuan Xu and Kate Saenko. 2016. Ask, Attend and Answer: Exploring Question-guided Spatial Attention for Visual Question Answering. In *ECCV*. 451–466.

[41] Fei Yan and Krystian Mikolajczyk. 2015. Deep Correlation for Matching Images and Text. In *CVPR*. 3441–3450.

[42] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. 2017. Pairwise Relationship Guided Deep Hashing for Cross-Modal Retrieval. In *AAAI*. 1618–1625.

[43] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked Attention Networks for Image Question Answering. In *CVPR*. 3441–3450.

[44] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *TACL* 2 (2014), 67–78.

[45] Deming Zhai, Xianming Liu, Hong Chang, Yi Zhen, Xilin Chen, Maozu Guo, and Wen Gao. 2018. Parametric Local Multiview Hamming Distance Metric Learning. *Pattern Recognition* 75 (2018), 250–262.

[46] Deming Zhai, Xianming Liu, Xiangyang Ji, Shin'ichi Satoh, Debin Zhao, and Wen Gao. 2018. Supervised Distributed Hashing for Large-scale Multimedia Retrieval. *IEEE TMM* 20, 3 (2018), 675–686.

[47] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2013. Heterogeneous Metric Learning with Joint Graph Regularization for Cross-media Retrieval. In *AAAI*. 1198–1204.

[48] C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. 2013. Learning the Visual Interpretation of Sentences. In *ICCV*. 1681–1688.