Image and Sentence Matching via Semantic Concepts and Order Learning

Yan Huang, Qi Wu, Wei Wang, and Liang Wang, *Fellow, IEEE*

Abstract—Image and sentence matching has made great progress recently, but it remains challenging due to the existing large visual-semantic discrepancy. This mainly arises from two aspects: 1) images consist of unstructured content which is not semantically abstract as the words in the sentences, so they are not directly comparable, and 2) arranging semantic concepts in different semantic order could lead to quite diverse meanings. The words in the sentences are sequentially arranged in a grammatical manner, while the semantic concepts in the images are usually unorganized. In this work, we propose a semantic concepts and order learning framework for image and sentence matching, which can improve the image representation by first predicting semantic concepts and then organizing them in a correct semantic order. Given an image, we first use a multi-regional multi-label CNN to predict its included semantic concepts in terms of object, property and action. These word-level semantic concepts are directly comparable with the words of noun, adjective and verb in the matched sentence. Then, to organize these concepts and make them express similar meanings as the matched sentence, we use a context-modulated attentional LSTM to learn the semantic order. It regards the predicted semantic concepts and image global scene as context at each timestep, and selectively attends to concept-related image regions by referring to the context in a sequential order. To further enhance the semantic order, we perform additional sentence generation on the image representation, by using the groundtruth order in the matched sentence as supervision. After obtaining the improved image representation, we learn the sentence representation with a conventional LSTM, and then jointly perform image and sentence matching and sentence generation for model learning. Extensive experiments demonstrate the effectiveness of our learned semantic concepts and order, by achieving the state-of-the-art results on two public benchmark datasets.

Index Terms—Semantic concept, semantic order, context-modulated attention, image and sentence matching

1 Introduction

The task of image and sentence matching refers to measuring the visual-semantic similarity between an image and a sentence. It has been widely applied to the application of image-sentence cross-modal retrieval, e.g., given an image query to find similar sentences, namely image annotation, and given a sentence query to retrieve matched images, namely text-based image search. Recently, much progress in this field has been achieved, but it is still a non-trivial task due to the intrinsic huge visual-semantic discrepancy. The discrepancy is mainly affected by two latent factors, namely semantic concept and semantic order.

Taking an image and its matched sentence in Fig. 1 for example, main objects, properties and actions appearing in the image are: {*cheetah*, *gazelle*, *grass*}, {*quick*, *young*, *green*} and {*chasing*, *running*}, respectively. These high-level semantic concepts are the essential content to be compared with the matched sentence, but they cannot be easily represented from the pixel-level image. Most existing methods [21], [26], [33] indistinctively represent all the concepts together by extracting the image global scene, in which the concepts are tangled with each other. As a result, some primary concepts in the foreground tend to be dominant, while other secondary background ones will probably be ignored, which is not optimal for fine-grained image and sentence matching. To comprehensively predict all the concepts for the image, a possible approach is to adapt the attribute learning frameworks [12], [55], [56] for concept prediction. But such a approach has not been well investigated in the context of image and sentence matching.

In addition to semantic concepts, how to correctly organize them, namely semantic order, plays an even more important role in the visual-semantic discrepancy. As illustrated in Fig. 1, given the semantic concepts mentioned above, if we incorrectly set their semantic order as: *a quick gazelle is chasing a young cheetah on grass*, then it would have completely different meanings compared with the image content, as well as the matched sentence. But directly learning the correct semantic order from separated semantic concepts is very difficult, since there potentially exist too many incorrect orders that could semantically make sense. We could resort to the image global

- *Y. Huang is with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100864, China, and with the University of Chinese Academy of Sciences (UCAS), Beijing 100049, P.R. China. E-mail: yhuang@nlpr.ia.ac.cn.*
- *Q. Wu is with the Australia Centre for Robotic Vision (ACRV), University of Adelaide, Adelaide, SA 5005, Australia. E-mail: qi.wu01@adelaide.edu.au.*
- *W. Wang and L. Wang are with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100864, China, and with the University of Chinese Academy of Sciences (UCAS), Beijing 100049, P.R. China. E-mail: {wangwei, wangliang}@nlpr.ia.ac.cn.*

Manuscript received 1 July 2018; revised 11 Oct. 2018; accepted 11 Oct. 2018. Date of publication 28 Nov. 2018; date of current version 4 Feb. 2020.
(Corresponding author: Yan Huang.)
Recommended for acceptance by T. Berg.
Digital Object Identifier no. 10.1109/TPAMI.2018.2883466

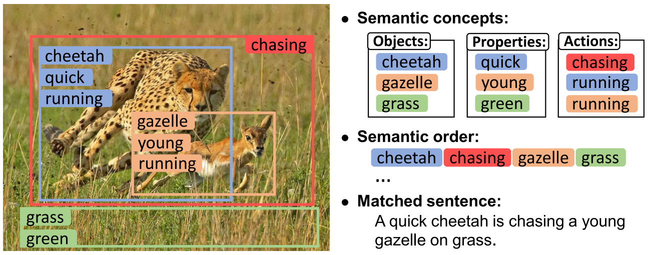


Fig. 1. Illustration of the semantic concepts and semantic order (best viewed in colors).

scene as auxiliary information, which indicates the correct semantic order as spatial configurations among semantic concepts, e.g., the cheetah is on the left of the gazelle. But it is unclear how to effectively use them to organize the semantic concepts, and make the learnt semantic order be directly comparable with the sequential order in the matched sentence.

Alternatively, we could generate a descriptive sentence from the image (i.e., to image captioning) as its textual representation, which learns both semantic concepts and order in an end-to-end manner. Thus the image and sentence matching problem is transformed to the task of matching the generated sentence and matched one. However, the image captioning itself is already a very challenging problem, even those state-of-the-art image captioning methods [49] cannot always generate very realistic sentences that capture all image concepts. Such information loss will inevitably degenerate the quality of the following sentence matching. In fact, these image captioning methods usually cannot achieve high performance for image and sentence matching [9], [49]. To reduce the information loss, we also make the attempt to directly sort already extracted words of semantic concepts, but find it still cannot achieve well performance due to the limited number of concepts, as explained in Section 3.3.

In this work, to bridge the visual-semantic discrepancy between image and sentence, we propose a semantic concepts and order learning framework, which improves the image representation by first learning semantic concepts and then organizing them in a correct semantic order. To learn the semantic concepts, we exploit a multi-regional multi-label convolutional neural network (CNN) that can simultaneously predict multiple concepts in terms of object, property, action, etc. The input of this CNN contains multiple selectively extracted proposals from the image, which can comprehensively capture all the semantic concepts regardless of whether they are in the foreground or background.

To organize the extracted semantic concepts in a semantic order, we develop a context-modulated attentional LSTM. The LSTM can selectively focus on image regions related to different concepts at each timestep, recurrently shift its attention by referring to image context across adjacent timesteps, and sequentially aggregate representations of concept-related regions within several timesteps in a sequential order. The image context is the gated fusion of semantic concepts and image global scene, which includes not only individual concepts but also their spatial configurations to guide the semantic order learning. To further enhance the semantic order, we use the groundtruth order in the matched sentence as supervision and force the aggregated image representation to be able to generate the matched sentence.

After improving the image representation with both semantic concepts and order learning, we match it with the sentence representation obtained by a conventional LSTM [19]. Then the whole model are learnt by jointly performing image and sentence matching and sentence generation, with a structured objective and a generation objective, respectively. To demonstrate the effectiveness of the proposed model, we perform experiments of image annotation and retrieval on two publicly available datasets: Flickr30K [60] and MSCOCO [30], and achieve the state-of-the-art results.

Our main contributions can be summarized as follows.

- We propose a semantic-enhanced image and sentence matching framework, where semantic concepts and order can be effectively learnt by multi-regional multi-label CNN and context-modulated attentional LSTM, respectively.
- We model the context modulation in the attentional procedure, which explicitly incorporates semantic concepts and global scene as reference information for accurate localization of concept-related regions.
- We demonstrate the joint learning of image and sentence matching and sentence generation can greatly benefit both two individual tasks.
- We achieve the current state-of-the-art performance on image and sentence matching .

It should be noted that this paper is a systematic extension of our preliminary conference versions [21], [22]. The present work adds to the initial versions in significant ways. First, we introduce the explicit semantic order learning with context-modulated attentional LSTM on concept-related regions (in Section 3.3.1). Experimentally, we demonstrate that performance can be promoted in comparison to the previous implicit order learning that purely focuses on concepts. Second, we improve the context-modulated attention by incorporating a gated fusion scheme on semantic concepts and global scene as image context (in Section 3.3.2). Third, we add considerable new experimental results in terms of ablation study, parameter analysis, and intuitive evaluation. In addition, we compare with a number of recently published papers and confirm that our model still outperforms existing methods using multiple evaluation metrics.

2 Related Work

2.1 Visual-Semantic Embedding Based Methods

Frome et al. [13] propose the first visual-semantic embedding framework, in which ranking loss, CNN [28] and Skip-Gram [35] are used as the objective, image and word encoders, respectively. Under the similar framework, Kiros et al. [25] replace the Skip-Gram with LSTM [19] for sentence representation learning, Vendrov et al. [47] use a new objective that can preserve the order structure of visual-semantic hierarchy, and Wang et al. [51] additionally consider within-view constraints to learn structure-preserving representations. Yan and Mikolajczyk [58] associate the image and sentence using deep canonical correlation analysis as the objective, where the matched image-sentence pairs have high correlation. Based on the similar framework, Klein et al. [26] use Fisher Vectors (FV) [39] to learn more discriminative representations for sentences, and

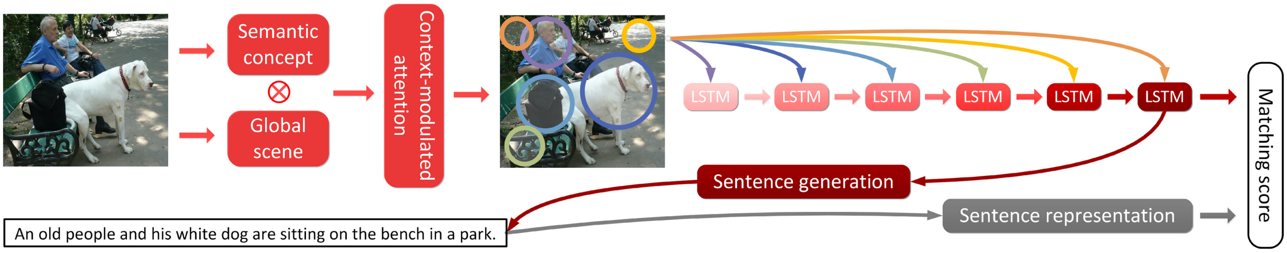


Fig. 2. The proposed semantic concepts and order learning framework (best viewed in colors).

Lev et al. [29] alternatively use RNN to aggregate FV and further improve the performance.

In addition to the global matching methods, Karpathy et al. [23], [24] make the first attempt to perform local similarity learning between fragments of images and sentences with a structured objective. Plummer et al. [40] collect region-to-phrase correspondences for instance-level image and sentence matching. Ma et al. [33] exploit a multimodal CNN for matching image and sentence, which can be regarded as an end-to-end framework for similarity score prediction. In contrast to them, we consider to extract semantic concepts from local image regions and then learn their semantic order.

2.2 Image Captioning Based Methods

Image captioning methods [12] can also be extended to deal with image-sentence matching, by first generating the sentence given an image and then comparing the generated sentence with groundtruth one. Chen and Zitnick [7] use a multimodal auto-encoder for bidirectional mapping, and measure the similarity using the cross-modal likelihood and reconstruction error. Mao et al. [34] propose a multimodal RNN model to generate sentences from images, in which the perplexity of generating a sentence is used as the similarity. Donahue et al. [9] design a long-term recurrent convolutional network for image captioning, which is also extended to image and sentence matching as well. Vinyals et al. [48], [49] develop a neural image captioning generator and show the effectiveness on the image and sentence matching. However, these models are originally designed to predict grammatically-complete sentences, so their performance on measuring the image-sentence similarity is not very well. Different from them, our work focuses on the similarity measurement, which is especially suitable for the task of image and sentence matching.

2.3 Deep Attention Based Methods

Our proposed model is also related to some methods simulating visual attention [52]. Alex Graves [15] exploits RNNs and differentiable Gaussian filters to simulate the attention mechanism, and applies it to handwriting synthesis. Gregor et al. [16] introduce the deep recurrent attentive writer for image generation, which develops a novel spatial attention mechanism based on 2-dimensional Gaussian filters to mimic the foveation of human eyes. Ba et al. [3] present a recurrent attention model that can attend to some label-relevant image regions of an image for multiple objects recognition. Bahdanau et al. [4] propose a neural machine translator which can search for relevant parts of a source sentence to predict a target word. Xu et al. [57] develop an attention-based model which can automatically learn to fix gazes on salient objects in an image and generate the corresponding annotated words. Different from these models, this work focuses more on the modelling of context information [1] during attention to compensate for the lack of semantic information, and propose context-modulated attention to find concept-related image regions for semantic order learning.

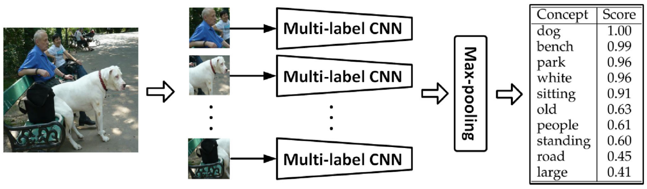
3 Semantic Concepts and Order Learning

We illustrate our proposed semantic concepts and order learning framework for image and sentence matching in Fig. 2. Given an image, we first predict its semantic concepts, e.g., *people* and *dog*, from local image regions, as well as the global scene indicating configurations among these concepts. Then the predicted concepts and scene are adaptively combined in a gated way, which serve as the reference information in the following context-modulated attention module to find concept-related image regions (marked by circles with different colors). To learn the semantic order of these concepts, we sequentially feed the concept-related regions into a LSTM at different timesteps, and then aggregate them in a desired semantic order. Note that the semantic order is not only learnt during the cross-modal matching with the representation of matched sentence, but also enhanced from an additional procedure of sentence generation under the supervision of groundtruth order in the matched sentence.

In the next, we will detail the proposed semantic concepts and order learning framework from the following aspects: 1) sentence representation learning with a conventional LSTM, 2) semantic concept extraction with a multi-regional multi-label CNN, 3) semantic order learning with a context-modulated attentional LSTM and sentence generation, and 4) model learning with joint image and sentence matching and sentence generation.

3.1 Sentence Representation Learning

For sentences, their included words of noun, adjective and verb directly correspond to the visual semantic concepts of object, property and action, respectively. The semantic order of these semantic-related words is intrinsically exhibited by the sequential nature of sentence. To learn the sentence representation that can capture those semantic-related words and model their semantic order, we use a conventional LSTM, similar to [25], [47]. The LSTM has multiple components for information memorizing and forgetting, which can well suit the complex properties of semantic concepts and order. We sequentially feed all the words of a sentence into the LSTM at different timesteps, and then regard the hidden state at the last timestep as the desired sentence representation $\mathbf{s} \in \mathbb{R}^H$.



Concept	Score
dog	1.00
bench	0.99
park	0.96
white	0.96
sitting	0.91
old	0.63
people	0.61
standing	0.60
road	0.45
large	0.41

Fig. 3. The multi-regional multi-label CNN for image concept extraction.

3.2 Image Semantic Concept Extraction

For images, their semantic concepts refer to various objects, properties and actions, which are represented in the visual content. The existing datasets do not provide this information for images at all but only matched sentences, so we have to predict them with an additional model. To learn such a model, we manually build a training dataset following [12], [55]. In particular, we only keep the nouns, adjectives, verbs and numbers as semantic concepts, and eliminate all the semantic-irrelevant words from the sentences. Considering that the size of concept vocabulary could be very large, we ignore those words that have very low use frequencies. In addition, we unify the different tenses of verb, and the singular and plural forms of noun to further reduce the vocabulary size. Finally, we obtain a vocabulary containing K semantic concepts. Based on this vocabulary, we can generate the training dataset by selecting multiple words from sentences as the groundtruth semantic concepts.

Then, the prediction of semantic concepts is equivalent to a multi-label classification problem. Many effective models on this problem have been proposed recently [14], [50], [53], [54], [55], which mostly learn various CNN-based models as nonlinear mappings from images to the desired multiple labels. Similar to [53], [55], we use the VGGNet [44] pretrained on the ImageNet dataset [42] as our multi-label CNN. To suit the multi-label classification, we modify the output layer to have K outputs, each corresponding to the predicted confidence score of a semantic concept. We then use the sigmoid activation instead of softmax on the outputs, so that the task of multi-label classification is transformed to multiple tasks of binary classification [20]. Given an image, its multi-hot representation of groundtruth semantic concepts is $\mathbf{y}_i \in \{0,1\}^K$ and the predicted score vector by the multi-label CNN is $\mathbf{p}_i \in [0,1]^K$, then the model can be learned by optimizing the following objective:

$$L_{cnn} = \sum_{c=1}^{K} \log\left(1 + e^{(-\mathbf{y}_{i,c}\mathbf{p}_{i,c})}\right). \tag{1}$$

The optimization can be regarded as a fine-tuning process, in which Dropout [45] is used to reduce the over-fitting.

During testing, we perform the concept prediction in a regional way, because the semantic concepts usually vary in size and appear in different locations including both foreground and background. As shown in Fig. 3, given a testing image, we first extract hundreds of region proposals using Multiscale Combinatorial Grouping (MCG) [41], and then cluster them into c clusters of hypothesis using Normalized Cut (NCut) [43]. By selecting the top h hypotheses from each cluster according to their predictive scores, we obtain totally $c \times h$ image regions, similar to [53]. By resizing these regions to square shapes and separately feeding them into the learned multi-label CNN, we can obtain a set of predicted confidence score vectors of local semantic concepts. We then perform element-wise max-pooling across these score vectors to obtain a single vector, which includes the desired confidence scores for all the semantic concepts.

3.3 Image Semantic Order Learning

After obtaining the semantic concepts, how to reasonably organize them in a correct semantic order plays an essential role to the image and sentence matching. Even though based on the same set of semantic concepts, combining them in different orders could lead to completely opposite meanings. For example in Fig. 3, if we organize the extracted semantic concepts including *dog*, *sitting* and *people* as: *a people is sitting on the dog*, then its meaning would be very different from the original image content.

To learn the desired order, a straightforward way is to select the top-k extracted concepts according to the predicted scores, and then sort them in the word level. But the size of concept vocabulary is very limited, many excluded concepts with low use frequencies (e.g., *otter*) are usually predicted as appearance similar but incorrect concepts (e.g., *dog*). The incorrect concepts are inconsistent with the image content, which cannot be well associated with the sentence and thus degenerate the performance. Therefore, we do not directly select and sort the concepts in the word level, but instead in the level of concept-related image regions. We actually use the semantic concepts in a more "soft" manner, by correlating them with the original image to find related regions. In this way, the incorrectly predicted concepts can also find the right image regions due to their similar appearances, so that the model can largely tolerate the potential errors.

3.3.1 *Context-Modulated Attention*

For an image as shown in Fig. 4a, directly obtaining its concept-related regions is difficult, since the image content is very complex where the semantic concepts could appear in any location with various scales. Considering that 1) not all image regions are necessary since images consist of too much concept-irrelevant information, and 2) the desired concept-related regions usually exist as a combination of multiple evenly divided boxes, e.g., the concept of *dog* covers about twelve boxes, we decide to first predict the concept-related attention weights for all the boxes to highlight those important ones, and then fuse their representations according to their importance to finally represent the desired concept-related regions.

To achieve this goal, we develop the context-modulated attentional LSTM, which selectively attends to multiple image regions by predicting a sequence of concept-related attention maps. It then explicitly aggregates their representations in the sequential manner of LSTM, in which the sequential order can be regarded as the desired semantic order for concepts. Different from traditional attentional LSTM [57], here we systematically study the roles of semantic concepts and image global scene in the attentional procedure for accurately uncovering the concept-related regions. It results from a motivation that each semantic concept seldom occurs in isolation but co-varies with other ones under a particular scene. There are also evidences from neuroscience showing that the global scene enables humans to quickly guide their

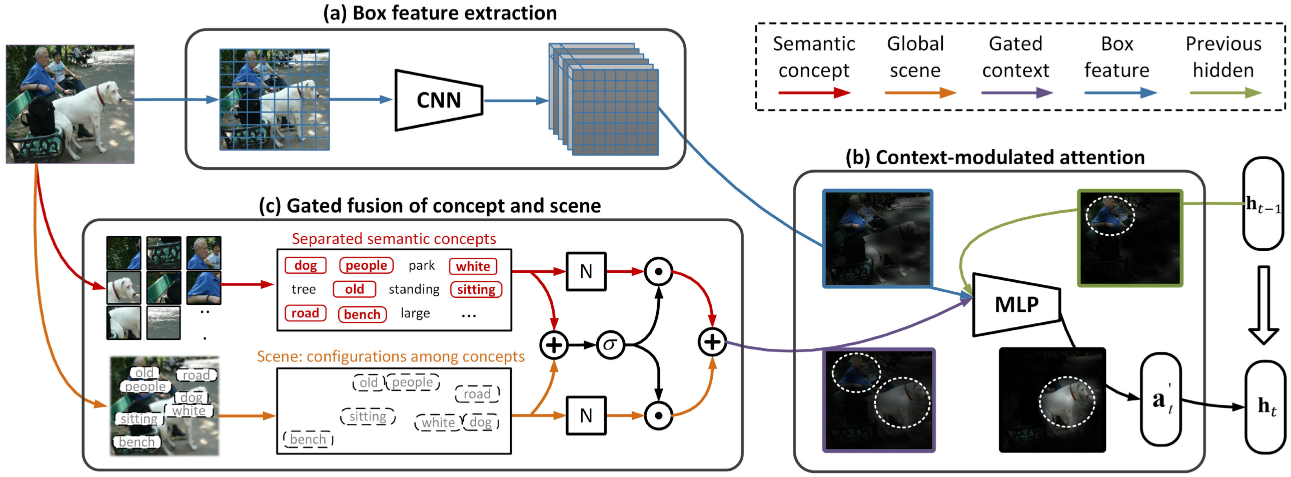


Fig. 4. Details of the proposed context-modulated attentional LSTM including a) box feature extraction, b) context-modulated attention, and c) gated fusion of concepts and scene (best viewed in colors).

attention to regions of interest [37], e.g., the concept-related regions in our case.

As illustrated in Fig. 4a, we represent the evenly divided boxes by extracting feature maps of the last convolutional layer in a CNN. We concatenate feature values at the same location across different feature maps as the feature vector for the corresponding convolved region, e.g., the concatenated vector in the top left of feature maps. We denote the feature set as $\{\mathbf{a}_i | \mathbf{a}_i \in \mathbb{R}^F\}_{i=1,\ldots,I}$, where $\mathbf{a}_i$ is the representation of the ith divided box, I is the total number of boxes, and F is the number of feature maps. We denote the previously extracted score vector of semantic concepts as $\mathbf{p} \in \mathbb{R}^K$, and the global scene for the image as $\mathbf{x} \in \mathbb{R}^D$. Based on these variables, we can perform concept-related attention map prediction at the tth timestep as follows:

$$p_{t,i} = e^{\hat{p}_{t,i}} \Big/ \sum_{i=1}^{I} .e^{\hat{p}_{t,i}},\ \ \hat{p}_{t,i} = f(\mathbf{x}, \mathbf{p}, \mathbf{a}_i, \mathbf{h}_{t-1}), \qquad (2)$$

where $p_{t,i}$ is the attention weight indicating the probability that the ith box will be attended to at the tth timestep, and $\mathbf{h}_{t-1}$ is the hidden state at the previous timestep. $f(\cdot)$ is a MultiLayer Perceptron (MLP) based function implementing the context-modulated attention, which is explained as follows.

During the attentional procedure in Fig. 4b, the independently learnt representations of divided boxes (marked by blue arrows) are used to compute the initial attention map, which has little information of semantic concepts. The hidden state at the previous timestep (marked by green arrows) indicates the already attended concept-related regions in the image, e.g., “man”. To select which concept-related region to attend to next, the attention scheme should first refer to all the concept candidates, namely image context (marked by purple arrows), to find a concept, and then compare it with previous hidden state to see if this concept has already been attended to. If yes (e.g., selecting the “man”), the scheme will refer to the image context again to find another concept. Otherwise (e.g., selecting the “dog”), regions in the initial attention map corresponding to the concept will be highlighted.

In such a context-modulated attentional procedure, the information in initial attention map is additively modulated by the image context and subtractively modulated by the previous hidden state, to finally produce the concept-related attention map. To efficiently simulate this, we use a three-way MLP as follows:

$$\begin{aligned} f(\mathbf{x}, \mathbf{p}, \mathbf{a}_i, \mathbf{h}_{t-1}) = \mathbf{w}(\sigma(g(\mathbf{x}, \mathbf{p})W_g + \mathbf{b_v}) + \sigma(\mathbf{a}_i W_\mathbf{a} + \mathbf{b_a}) \\ + \sigma(\mathbf{h}_{t-1} W_\mathbf{h} + \mathbf{b_h})) + b, \end{aligned} \qquad (3)$$

where σ denotes the sigmoid activation function, $g(\cdot)$ is the gated function to compute the image context, and $\mathbf{w}$ and b are a weight vector and a scalar bias, respectively. W_g, $W_\mathbf{a}$ and $W_\mathbf{h}$ are weight matrices associated with the image context, box representations and hidden state, respectively. Note that the mentioned initial attention map is implicitly computed by using the $\sigma(\mathbf{a}_i W_\mathbf{a} + \mathbf{b_a})$ term in this equation to obtain attention weights.

3.3.2 Gated Fusion of Concepts and Scene as Context

In the equation above, the image context $g(\mathbf{x}, \mathbf{p})$ is a gated combination of semantic concepts $\mathbf{p}$ and global scene $\mathbf{x}$. There are two main reasons accounting for introducing the global scene as follows. 1) It is uneasy to decide the semantic order only from separated semantic concepts during attention, since the order involves not only the hypernym relations between concepts, but also the textual entailment among phrases in high levels of semantic hierarchy [47]. And 2) as illustrated in Fig. 4c, the global scene can not only describe all the semantic concepts in a coarse level, but also indicate their spatial configurations with each other, e.g., a people is sitting on the bench in the left while the dog in the middle right. Such spatial configurations have been demonstrated can cause the perception of one concept to generate strong expectations about other concepts [8], [37], which can help our model to guide its attention from one observed concept to an unobserved one, i.e., ordering the concepts. Therefore, we propose to use the global scene[1] as auxiliary reference to facilitate the semantic order learning.

[1] In practice, for efficient implementation, we use a pre-trained VGGNet to process the whole image content, and then extract the vector in the last fully-connected layer as the desired global scene.

To model such a reference procedure, a simple way is to equally sum the global scene with semantic concepts together as image context. But considering that the content of different images can be diverse, so the relative importance of semantic concepts and scene is not equivalent in most cases. For those images with complex content, the predictions of their concepts might be distracted by background, so the global scene might be more reliable. Inspired by this, we design a gated fusion unit that can selectively balance the relative importance of semantic concepts and scene. The unit acts as a gate that controls how much information of the semantic concepts and scene contributes to their fused image context. As illustrated in Fig. 4c, after obtaining the global scene vector $\mathbf{x}$ and concept score vector $\mathbf{p}$, their gated fusion can be formulated as:

$$\begin{aligned}&\widehat{\mathbf{x}} = \|W_x\mathbf{x}\|_2,\ \widehat{\mathbf{p}} = \|W_p\mathbf{p}\|_2,\ \mathbf{t} = \sigma(U_x\mathbf{x} + U_p\mathbf{p})\\ &g(\mathbf{x},\mathbf{p}) = \mathbf{t}\odot\widehat{\mathbf{x}} + (\mathbf{1}-\mathbf{t})\odot\widehat{\mathbf{p}},\end{aligned} \tag{4}$$

where $\|\cdot\|_2$ denotes the l_2-normalization. The use of sigmoid function σ is to rescale each element in the gate vector $\mathbf{t}\in\mathbb{R}^H$ to $[0,1]$, so that $g(\mathbf{x},\mathbf{p})$ becomes an element-wise weighted sum of $\mathbf{x}$ and $\mathbf{p}$. Different from [38], our gated fusion unit especially uses the l_2-normalization during gating to re-scale the transformed concepts and context, so that the fused vector is robust to the cross-modal similarity computation. If the normalization is eliminated, we experimentally find the performance becomes even worse than the simple summation.

3.3.3 Concept-Related Region Aggregation

According to the predicted concept-related attention maps by context-modulated attention, we compute the weighted sum representations $\mathbf{a}'_t$ to adaptively describe the attended image regions. We sum all the products of element-wise multiplication between each box representation (e.g., $\mathbf{a}_i$) and its corresponding attention weight (e.g., $p_{t,i}$):

$$\mathbf{a}'_t = \sum_{i=1}^{I} p_{t,i}\mathbf{a}_i, \tag{5}$$

where image boxes with higher attention weights contribute more to the representations of concept-related region.

From the 1st to Tth timestep, we obtain a sequence of representations of concept-related regions $\{\mathbf{a}'_t\}_{t=1,\cdots,T}$, where T is the total number of timesteps. To aggregate these regions for the whole image representation, we use a LSTM network to sequentially take them as inputs, where the hidden states $\{\mathbf{h}_t\in\mathbb{R}^H\}_{t=1,\ldots,T}$ dynamically propagate the representations of image regions until the end. The LSTM includes various gated mechanisms which can well suit the complex nature of semantic order. The hidden state at the last timestep $\mathbf{h}_T$ can be regarded as the desired image representation with semantic order.

Note that the attentional LSTMs in [3], [4], [57] do not consider the modeling of context modulation in their attention schemes, so they have to alternatively use step-wise labels to guide the prediction of attended regions. But such strong supervision can only be available for limited tasks, e.g., the sequential words of sentence for image captioning [57], and multiple class labels for multi-object recognition [3]. In fact, we perform experiments without using the context modulation in Section 4.7, but find that some concept-related regions like "cat" and "giraffe" cannot be well attended to. It mainly results from that the attention scheme can only refer to the initial attention map to select which concept to attend to next, but the initial attention map is computed from box representations which contains little information of semantic concepts and order.

3.3.4 Sentence Generation as Supervision

To learn the semantic concepts and order for image and sentence matching, an possible end-to-end approach is to generate a sentence directly from the image, similar to image captioning [55]. By regarding the predicted sentence as the representation of image, the task of image and sentence matching is thus transformed to the matching task between generated sentence and groundtruth one. However, the performance of such an approach heavily relies on the quality of generated sentence. In particular, although the current image captioning methods can generate semantically meaningful sentences with desired order, the accuracy of their generated sentences on capturing image concepts is not very high. Even a small error in the sentences can be amplified and affects the following similarity measurement in a fine-grained word level. Actually, even the state-of-the-art image captioning models [9], [34], [49] cannot perform very well on the image and sentence matching task. We also implement a similar model (as "sce + sen") for comparison in Section 4.3, and find it only achieves inferior results.

In fact, we do not have to generate a grammatically-complete sentence as our image representation, since we already have the aggregated image representation from concept-related regions. But we can enhance the semantic order learning by supervising the image representation with groundtruth semantic order in the matched sentence during a sentence generation precedure, just like image captioning. In particular, we feed the image representation into the initial hidden state of a generative LSTM, and ask it to be capable of generating the matched sentence. During the cross-word and cross-phrase generations, the image representation can thus learn the hypernym relations between words and textual entailment among phrases as the semantic order. Given a sentence $\{\mathbf{w}_j|\mathbf{w}_j\in\{0,1\}^G\}_{j=1,\ldots,J}$, where each word $\mathbf{w}_j$ is represented as an one-hot vector, J is the length of the sentence, and G is the size of word dictionary, we can formulate the sentence generation as follows:

$$\begin{aligned}&\mathbf{i}_t = \sigma(W_{\mathbf{wi}}(F\mathbf{w}_t) + W_{\mathbf{hi}}\mathbf{h}_{t-1} + \mathbf{b_i}),\\ &\mathbf{f}_t = \sigma(W_{\mathbf{wf}}(F\mathbf{w}_t) + W_{\mathbf{hf}}\mathbf{h}_{t-1} + \mathbf{b_f}),\\ &\mathbf{o}_t = \sigma(W_{\mathbf{wo}}(F\mathbf{w}_t) + W_{\mathbf{ho}}\mathbf{h}_{t-1} + \mathbf{b_o}),\\ &\widehat{\mathbf{c}}_t = \tanh(W_{\mathbf{wc}}(F\mathbf{w}_t) + W_{\mathbf{hc}}\mathbf{h}_{t-1} + \mathbf{b_c}),\\ &\mathbf{c}_t = \mathbf{f}_t\odot\mathbf{c}_{t-1} + \mathbf{i}_t\odot\widehat{\mathbf{c}}_t,\ \mathbf{h}_t = \mathbf{o}_t\odot\tanh(\mathbf{c}_t),\\ &\mathbf{q}_t = softmax(F^T\mathbf{h}_t + \mathbf{b}_p),\ e = \arg\max(\mathbf{w}_t),\\ &P(\mathbf{w}_t|\mathbf{w}_{t-1},\mathbf{w}_{t-2},\ldots,\mathbf{w}_0,\mathbf{x},\mathbf{p}) = \mathbf{q}_{t,e}\end{aligned} \tag{6}$$

where $\mathbf{c}_t$, $\mathbf{h}_t$, $\mathbf{i}_t$, $\mathbf{f}_t$ and $\mathbf{o}_t$ are memory state, hidden state, input gate, forget gate and output gate, respectively, e is the index of $\mathbf{w}_t$ in the word vocabulary, and $F\in\mathbb{R}^{D\times G}$ is a word embedding matrix. During the sentence generation, since all the words are predicted in a chain manner, the probability P of current predicted word is conditioned on all its previous words, as well as the input semantic concepts $\mathbf{p}$ and global scene $\mathbf{x}$ at the initial timestep.

3.4 Joint Matching and Generation

During the model learning, we jointly perform image and sentence matching and sentence generation, by minimizing the following combined objectives:

$$L = L_{mat} + \lambda \times L_{gen} + \mu \times L_{reg}, \tag{7}$$

where λ and μ are tuning parameters for balancing and regularization.

The L_{mat} is a structured objective that encourages the cosine similarity scores of matched images and sentences to be larger than those of mismatched ones:

$$\sum_{ik} \max\{0, m - s_{ii} + s_{ik}\} + \max\{0, m - s_{ii} + s_{ki}\},$$

where m is a margin parameter, s_{ii} is the score of matched ith image and ith sentence, s_{ik} is the score of mismatched ith image and kth sentence, and vice-versa with s_{ki}. We empirically set the total number of mismatched pairs for each matched pair as 128 in our experiments.

The L_{gen} is the negative conditional log-likelihood of the matched sentence given the semantic concepts $\mathbf{p}$ and global scene $\mathbf{x}$:

$$-\sum_{t} \log P(\mathbf{w}_t | \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \ldots, \mathbf{w}_0, \mathbf{x}, \mathbf{p}),$$

where the detailed formulation of probability P is shown in Equation (6). Note that we use the predicted semantic concepts rather than groundtruth ones in our experiments.

The L_{reg} is a regularization term [57] to constrain the sum of attention weights of any box at all timesteps as follows:

$$\sum_{i} \left(c - \sum_{t} p_{t,i} \right),$$

where c is a constant and we empirically find that setting $c = 1$ can lead to well performance. Without the regularization term, our model is inclined to focus on the same image regions at all timesteps. It might result from the fact that always selecting the most informative regions can largely avoid errors. But it is not good for our model to comprehensively perceive the entire image content to find different concept-related regions. So we add this term to encourage the model to pay equal attention to every box rather than a certain one for information maximization.

It should be noted that we do not need to generate the sentence during testing. We only have to compute the image representation and then compare it with the sentence representation $\mathbf{s}$ to obtain their similarity score.

4 Experimental Results

To demonstrate the effectiveness of the proposed model, we perform several experiments in terms of image annotation and retrieval on two publicly available datasets.

4.1 Datasets and Protocols

The two evaluation datasets and their experimental protocols are described as follows. 1) Flickr30k [60] consists of 31783 images collected from the Flickr website. Each image is accompanied with 5 human annotated sentences. We use the public training, validation and testing splits [25], which contain 28000, 1000 and 1000 images, respectively. 2) MSCOCO [30] consists of 82783 training and 40504 validation images, each of which is associated with 5 sentences. We use the public training, validation and testing splits [25], with 82783, 4000 and 1000 (or 5000) images, respectively. When using 1000 images for testing, we perform 5-fold cross-validation and report the averaged results.

4.2 Implementation Details

The commonly used evaluation criterions for image annotation and retrieval are "R@1", "R@5" and "R@10", i.e., recall rates at the top 1, 5 and 10 results. We also compute an additional criterion "mR" by averaging all the 6 recall rates, to evaluate the overall performance for both image annotation and retrieval.

We use the 19-layer VGGNet [44] to initialize the multi-regional multi-label CNN to predict semantic concepts. We also use the 19-layer VGG network to initialize another CNN to extract 512 feature maps (with a size of 14 × 14) in "conv5-4" layer as the representations for evenly divided boxes, and a feature vector in "fc7" layer as the image global scene. The dimensions of box representations and global scene features are $F = 512$ and $D = 4096$, respectively, and the total number of boxes is $I = 196$ (14 × 14). We perform 10-cropping [26] from the images and then separately feed the cropped regions into the network. The final global scene is averaged over 10 cropped regions. In addition, we also try to use the 152-layer ResNet [18] to initialize the CNN. We accordingly extract 2048 feature maps (with a size of 7 × 7) in the last convolutional layer as the representations for evenly divided boxes, and a feature vector in the last fully-connected layer as the image global scene. The dimensions of box representations and global scene features are $F = 2048$ and $D = 1000$, respectively, and the total number of boxes is $I = 49$ (7 × 7).

For sentences, the dimension of embedded word is $D = 300$. We set the max length for all the sentences, i.e., the number of words, as $J = 50$, and use zero-padding when a sentence is not long enough. Other parameters are empirically set including the number of clusters c=10, number of hypotheses $h = 5$, dimension of hidden state in LSTM $H = 1024$, number of semantic concepts $K = 256$, balancing parameter $\lambda = 1$, regularization parameter $\mu = 100$, and margin parameter $m = 0.2$.

During the model training, we use stochastic gradient descent with a learning rate of 0.01, momentum of 0.9, weight decay of 0.0005, batch size of 128, and gradient clipping at 0.1. The model is trained for 30 epochs to guarantee its convergence.

4.3 Study of Ablation Models

To systematically evaluate the contributions of different model components, we design various ablation models as shown in Table 1. The variable model components are explained as follows: 1) "sce (1-crop)" and "sce (10-crop)"

TABLE 1
Comparison Results of Image Annotation and Retrieval by Ablation Models on the Flickr30k and MSCOCO (1000 Testing) Datasets

Method	Flickr30k dataset							MSCOCO dataset						
	Image Annotation			Image Retrieval			mR	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
sce (1-crop)	29.8	58.4	70.5	22.0	47.9	59.3	48.0	43.3	75.7	85.8	31.0	66.7	79.9	63.8
sce (10-crop)	33.8	63.7	75.9	26.3	55.4	67.6	53.8	44.7	78.2	88.3	37.0	73.2	85.7	67.9
cnp (VG)	41.3	68.8	78.8	30.5	59.0	70.2	58.1	51.7	83.4	91.6	42.8	78.4	89.8	73.0
cnp	30.9	60.9	72.4	23.1	52.5	64.8	50.8	59.5	86.9	93.6	48.5	81.4	90.9	76.8
cnp + sce (C)	39.9	71.2	81.3	31.4	61.7	72.8	59.7	62.8	89.2	95.5	53.2	85.1	93.0	79.8
cnp + sce	42.4	72.9	81.5	32.4	63.5	73.9	61.1	65.3	90.0	96.0	54.2	85.9	93.5	80.8
iatt (cnps)	31.5	62.0	73.5	23.9	53.3	65.7	51.6	60.5	87.2	93.8	49.2	82.0	91.3	77.3
iatt	26.8	53.3	65.9	21.7	49.2	62.0	46.5	39.5	71.9	84.0	33.4	70.5	83.9	63.9
iatt + satt	27.0	53.4	65.5	20.5	49.7	61.6	46.3	40.1	72.6	84.4	32.6	69.5	83.2	63.7
iatt + cnp	33.6	63.1	74.5	24.4	53.6	65.8	52.5	62.7	89.5	94.9	51.1	83.4	92.0	78.9
iatt + sce	34.9	65.1	76.3	27.1	55.6	68.4	54.6	45.3	78.9	88.6	37.2	73.5	87.0	68.5
iatt + cnp + sce (R)	35.1	66.3	77.3	27.1	59.1	68.9	55.6	56.7	84.0	88.7	49.6	80.7	88.5	74.7
iatt + cnp + sce (B)	42.8	72.7	83.3	31.7	63.4	74.0	61.3	65.1	90.2	96.4	53.9	86.0	93.2	80.8
iatt + cnp + sce	43.1	73.4	83.1	32.9	63.9	74.3	61.8	66.2	90.9	96.5	54.5	86.4	93.8	81.4
sen + sce	22.8	48.6	60.8	19.1	46.0	59.7	42.8	39.2	73.3	85.5	32.4	70.1	83.7	64.0
gen + sce (S)	34.4	64.5	77.0	27.1	56.3	68.3	54.6	45.7	78.7	88.7	37.3	73.8	85.8	68.4
gen + sce (E)	35.5	63.8	75.9	27.4	55.9	67.6	54.3	46.9	78.8	89.2	37.3	73.9	85.9	68.7
gen + sce	35.6	66.3	76.9	27.9	56.8	68.2	55.3	46.9	79.2	89.3	37.9	74.0	85.9	68.9
gen + cnp	31.5	61.7	74.5	25.0	53.4	64.9	51.8	62.6	89.0	94.7	50.6	82.4	91.2	78.4
gen + cnp + sce	44.2	74.1	83.6	32.8	64.3	74.9	62.3	66.4	91.3	96.6	55.5	86.5	93.7	81.8
iatt + cnp + sce + gen	45.9	74.9	84.8	33.9	64.9	76.0	63.4	67.5	92.2	97.0	56.5	87.4	94.8	82.6

refer to using the global scene by cropping 1 or 10 regions from images, respectively. 2) “cnp” denotes using predicted semantic concepts, and “cnp (VG)” uses the external Visual Genome (VG) dataset [27] instead of our constructed one for semantic concept prediction. 3) “cnp + sce (C)” and “cnp + sce” are two different ways that combine semantic concepts and scene via direct summation and gated fusion unit, respectively. 4) “iatt” and “satt” perform the traditional attention on images and sentences, respectively. “iatt (cnps)” performs attention on multiple score vectors of semantic concept from local image regions rather than concept-related regions. “iatt + cnp” performs the proposed context-modulated attention by regarding “cnp” as its context. 5) “iatt + cnp + sce”, “iatt + cnp + sce (R)” and “iatt + cnp + sce (B)” organize the extracted concept-related regions with LSTMs in three different orders including forward, random and backward, respectively. The forward order learns to fuse attended regions along the time axis, random order performs the fusion by randomly selecting a region each time, and backward order reverses the learned forward order. 6) “sen + sce” uses the state-of-the-art image captioning method [49] to generate sentences from images and then regards the sentences as image representations for matching. Different from it, “gen + sce” enhances the semantic order by using the sentence generation as supervision as described in Section 3.3.4. 7) “gen + sce (S)” additionally uses the scheduled sampling [5]. “gen + sce (E)” indicates that the parameters of two word embedding matrices for sentence representation and sentence generation are shared.

The results of ablation models on the Flickr30k and MSCOCO datasets are shown in Table 1, from which we can obtain the following conclusions in three aspects.

Gated Fusion of Concepts and Scene. 1) Cropping 10 image regions (as “sce (10-crop)”) can achieve much robust global scene features than cropping only 1 region (as “sce (1-crop)”). 2) Only using the semantic concepts (as “cnp”) can already achieve good performance, especially when the training data are sufficient on the MSCOCO dataset. 3) Using the external VG dataset for semantic concept prediction (as “cnp (VG)”) can achieve better performance on the Flicker30k dataset but worse performance on the MSCOCO dataset. It is probably because the visual content of VG and MSOCOCO is quite different, and the learned predictor on one dataset cannot well generalize to the other one. 4) Simply summing the concepts and scene (as “cnp + sce (C)”) can further improve the result, because the image scene contains the spatial configurations of concepts that are complimentary to the semantic concepts. 5) Using the proposed gated fusion unit (as “cnp + sce”) performs better, due to the effective importance balancing scheme.

Context-Modulated Attention. 6) Performing the attention scheme on multiple local semantic concepts (as “iatt (cnps)”) can achieve better results than non-sorted concepts (as “cnp”), which demonstrates the effectiveness of semantic order learning. 7) Modeling concept-related image regions (as “iatt + cnp”) is more discriminative than just concepts (“iatt (cnps)”). 8) Only using convolutional features of image boxes (as “iatt”) gets much worse results, mainly due to the lack of high-level semantic concept information. 9) Additionally performing the attention on sentences (as “iatt + satt”) still cannot improve the performance, since the shuffled semantic order in sentences cannot benefit the order learning of images. 10) When regarding either concepts and scene as image context and performing context-modulated attention (as “iatt + cnp” or “iatt + sce”), the models can achieve much better performance. 11) By fusing them with the gated fusion unit (as “iatt + cnp + sce”), the performance can be further

TABLE 2
Comparison Results of Image Annotation and Retrieval on the Flickr30k and MSCOCO (1000 Testing) Datasets

Model	Flickr30k dataset							MSCOCO dataset						
	Image Annotation			Image Retrieval			mR	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
FV [26]	35.0	62.0	73.8	25.0	52.7	66.0	52.4	39.4	67.9	80.9	25.1	59.8	76.6	58.3
DVSA [24]	22.2	48.2	61.4	15.2	37.7	50.5	39.2	38.4	69.9	80.5	27.4	60.2	74.8	58.5
MNLM [25]	23.0	50.7	62.9	16.8	42.0	56.5	42.0	43.4	75.7	85.8	31.0	66.7	79.9	63.8
m-CNN [33]	33.6	64.1	74.9	26.2	56.3	69.6	54.1	42.8	73.1	84.1	32.6	68.6	82.8	64.0
RNN+FV [29]	34.7	62.7	72.6	26.2	55.1	69.2	53.4	40.8	71.9	83.2	29.6	64.8	80.5	61.8
OEM [47]	-	-	-	-	-	-	-	46.7	78.6	88.9	37.9	73.7	85.9	68.6
VQA [31]	33.9	62.5	74.5	24.9	52.6	64.8	52.2	50.5	80.1	89.7	37.0	70.9	82.9	68.5
RTP [40]	37.4	63.1	74.3	26.0	56.0	69.3	54.3	-	-	-	-	-	-	-
DSPE [51]	40.3	68.9	79.9	29.7	60.1	72.1	58.5	50.1	79.7	89.2	39.6	75.2	86.9	70.1
*sm-LSTM [21]	42.5	71.9	81.5	30.2	60.4	72.3	59.8	53.2	83.1	91.5	40.7	75.8	87.4	72.0
2WayNet [10]	**49.8**	67.5	-	**36.0**	55.6	-	-	55.8	75.2	-	39.7	63.3	-	-
DAN [36]	41.4	73.5	82.5	31.8	61.7	72.5	60.6	-	-	-	-	-	-	-
VSE++ [11]	41.3	69.0	77.9	31.4	59.7	71.2	58.4	57.2	85.1	93.3	45.9	78.9	89.1	74.6
*SCO [22]	44.2	74.1	83.6	32.8	64.3	74.9	62.3	66.6	91.8	96.6	55.5	86.6	93.8	81.8
Ours	45.9	**74.9**	**84.8**	33.9	**64.9**	**76.0**	**63.4**	**67.5**	**92.2**	**97.0**	**56.5**	**87.4**	**94.8**	**82.6**
RRF (Res) [32]	47.6	77.4	87.1	35.4	68.3	79.9	66.0	56.4	85.3	91.5	43.9	78.1	88.6	73.9
DAN (Res) [36]	55.0	81.8	89.0	39.4	69.2	79.1	68.9	-	-	-	-	-	-	-
VSE++ (Res) [11]	52.9	79.1	87.2	39.6	69.6	79.5	68.0	64.6	89.1	95.7	52.0	83.1	92.0	79.4
LIM (Res) [17]	-	-	-	-	-	-	-	68.5	-	97.9	56.6	-	94.5	-
BUTD (Res) [2]	53.1	81.9	88.9	40.1	69.8	79.7	68.9	65.6	91.8	96.9	54.6	85.8	93.3	81.3
*SCO (Res) [22]	55.5	82.0	89.3	41.1	70.5	80.1	69.7	69.9	92.9	97.5	56.7	87.5	94.8	83.2
Ours (Res)	**58.0**	**84.5**	**90.5**	**43.9**	**72.9**	**81.6**	**71.9**	**71.3**	**93.8**	**98.0**	**58.2**	**88.8**	**95.3**	**84.2**

* *Indicates our previous conference versions.*

improved, which verifies the effectiveness of context-modulated attention. 12) When either using a random semantic order (as "iatt + cnp + sce (R)") or reversing the semantic order from forward direction to back direction (as "iatt + cnp + sce (B)"), the performance degenerates heavily on both datasets. It demonstrates that our learned forward semantic order is more effective for image and sentence matching.

Sentence Generation as Supervision. 13) Directly using the pre-generated sentences as image representations (as "sce + sen") cannot improve the performance, since the generated sentences might not accurately include all the image concepts. 14) Using the sentence generation as supervision for semantic order learning (as "sce + gen") is very effective. 15) But additionally performing the scheduled sampling (as "sce + gen (S)") cannot further improve the performance. It is because the groundtruth semantic order is degenerated during sampling, accordingly the model cannot learn it well. 16) Using a shared word embedding matrix (as "sce + gen (E)") cannot improve the performance, which might result from that learning a unified matrix for two different tasks could be easily confused.

The best performance is finally achieved by the "iatt + cnp + sce + gen", which predicts the semantic concepts and then learns their semantic order via context-modulated attention and sentence generation. In the follow experiments, we regard the "iatt + cnp + sce + gen" as our default model.

4.4 Comparison with State-of-the-Art Methods

We compare our proposed model with recent state-of-the-art methods on the Flickr30k and MSCOCO datasets in Table 2, as well as our two preliminary conference versions: sm-LSMT [21] and SCO [22], marked by *. The methods marked by "(Res)" use the 152-layer ResNet [18] for scene extraction, while the rest ones use the default 19-layer VGGNet [44].

Using either VGGNet or ResNet on the MSCOCO dataset, our proposed model outperforms the current state-of-the-art models by a large margin on all 7 evaluation criterions. It demonstrates that learning semantic concepts and order for image representations is very effective. When using VGGNet on the Flickr30k dataset, our model gets lower performance than 2WayNet on the R@1 evaluation criterion, but obtains much better overall performance on the rest evaluation criterions. When using ResNet on the Flickr30k dataset, our model is able to achieve the best result. Note that our model obtains much larger improvements on the MSCOCO dataset than Flickr30k. It is because the MSCOCO dataset has more training data, so that our model can be better fitted to predict more accurate image-sentence similarities. Note that our current model achieves much better than our previous model SCO [22], which indicates that explicitly learning semantic order with context-modulated attentional LSTM on concept-related regions is more useful than previous implicit order learning purely on concepts. Our model also performs better than a recent approach BUTD [2], which first performs the conventional attention over outputs from a pretrained object detector, and then fuses all attended outputs to build the image representation. Especially on the Flickr30k dataset, even though the predicted semantic concepts from our constructed dataset are less discriminative than theirs from Visual Genome dataset (in Table 1), our overall performance is much better than theirs. It again demonstrates the effectiveness of our context-modulated attention and sentence generation for semantic order learning.

TABLE 3
Comparison Results of Image Annotation and Retrieval on the MSCOCO (5000 Testing) Dataset

Method	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
DVSA [24]	11.8	32.5	45.4	8.9	24.9	36.3	26.6
FV [26]	17.3	39.0	50.2	10.8	28.3	40.1	31.0
VQA [31]	23.5	50.7	63.6	16.7	40.5	53.8	41.5
OEM [47]	23.3	50.5	65.0	18.0	43.6	57.6	43.0
VSE++ [11]	32.9	61.6	74.7	24.1	52.0	66.2	51.9
*SCO [22]	40.2	70.1	81.3	31.3	61.5	73.9	59.7
Ours	**42.1**	**71.5**	**82.2**	**32.5**	**62.6**	**75.1**	**61.0**
VSE++ (Res) [11]	41.3	69.2	81.2	30.3	59.1	72.4	58.9
LIM (Res) [17]	42.0	-	84.7	31.7	-	74.6	-
BUTD (Res) [2]	41.2	71.4	82.8	32.2	62.0	74.5	60.7
*SCO (Res) [22]	42.8	72.3	83.0	33.1	62.9	75.5	61.6
Ours (Res)	**45.7**	**76.0**	**86.4**	**36.8**	**67.0**	**78.8**	**65.1**

* *Indicates our previous conference versions.*

The above experiments on the MSCOCO dataset follow the first protocol [24], which uses 1000 images and their associated sentences for testing. We also test the second protocol that uses all the 5000 images and their sentences for testing, and present the comparison results in Table 3. From the table we can observe that the overall results by all the methods are lower than the first protocol. It probably because the target set is much larger so there exist more distracters for a given query. Among all the models, the proposed model still achieves the best performance, which again demonstrates its effectiveness.

4.5 Evaluation of Predicted Semantic Concepts

Semantic Concept Visualization. To qualitatively verify the effectiveness of our multi-regional multi-label CNN for semantic concept prediction. We present several example images and their predicted semantic concepts in Fig. 6. From the figure we can see that the multi-regional multi-label CNN can predict reasonable semantic concepts with high confidence scores for describing the detailed image content. For example, *road*, *motorcycle* and *riding* are predicted from the second image. We also note that the *skate* is incorrectly assigned, which might result from that this image content is complicated and the smooth country road looks like some skating scenes.

Number of Semantic Concepts. To further study whether the number of semantic concepts has effects on the performance. We plot a curve in Fig. 7, in which the x-axis is the number of concepts and the y-axis is the averaged recall rate on the Flickr30k dataset. We can find that the performance changes when varying the number of concepts, and the larger number leads to better performance. Especially when reducing the number to 50, the performance drops heavily from 64 to 57 percent. It is mainly because that some images lack of appropriate concepts for description so that they cannot be well associated with sentences.

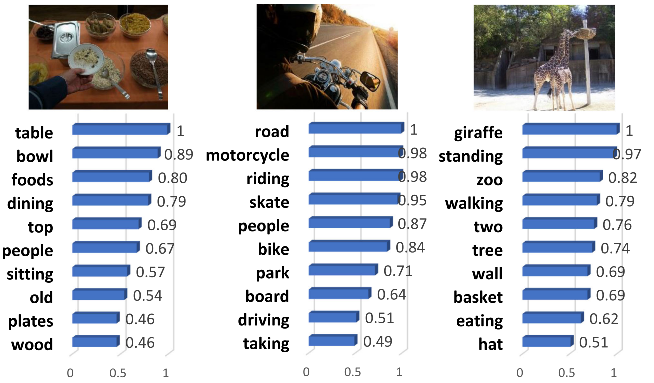


Fig. 6. Predicted top-10 semantic concepts with their confidence scores.

4.6 Evaluation of Learnt Semantic Order

Semantic Order Visualization. To verify whether the proposed model can selectively attend to concept-related image regions at different timesteps, as well as organize them in a semantic order, we visualize the predicted sequential attention maps by the proposed model in Fig. 5b. For each image, we resize the predicted attention weights at the tth timestep (with a size of 14×14) to the same size as its corresponding original image, so that each value in the resized map measures the importance of an image pixel at the same location. We then perform element-wise multiplication between the resized attention map and the original image to obtain the final attention map, where lighter areas indicate attended concept-related regions. From the figure we can see that our model can attend to image regions associated with different concepts at three timesteps. Taking the middle image for example, the model sequentially focuses on three highlighted regions indicating the concepts of "children", "playing" and "giraffe", respectively.

It seems that the model learns the semantic order by finding salient objects at the first and third timesteps, and their relations (e.g., actions or environments) in the second timestep. It is similar with most groundtruth semantic orders in the matched sentences, in which verbs are in the middle between two nouns. In Fig. 9, we also compute the averaged attention maps (rescaled to the same size of 500×500) for all the testing images at three different timesteps. We can see that the proposed model statistically tends to focus on the central regions at the first timestep, which is in consistent with the

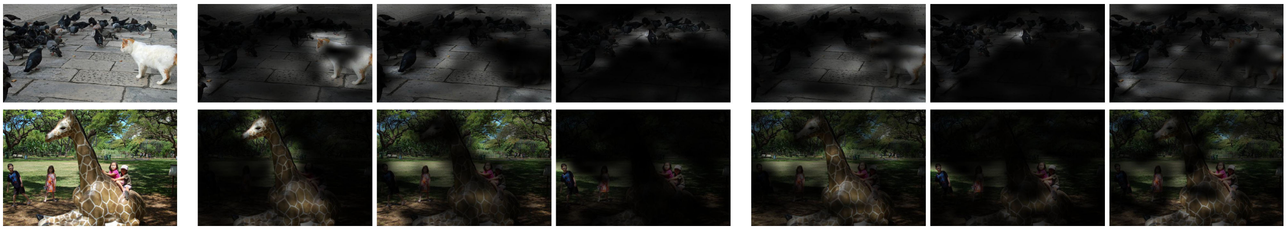

(a) Input image (b) Predicted attention maps using image context (c) Predicted attention maps without image context

Fig. 5. Input images and attended image regions at three different timesteps, using image context or not, respectively (best viewed in colors).

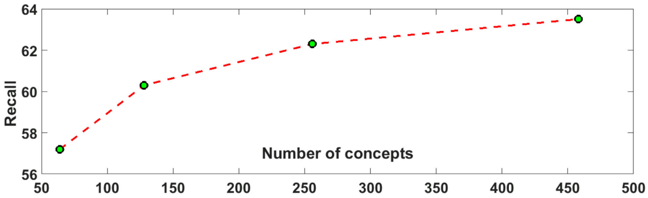


Fig. 7. Performance versus number of semantic concepts.

mechanism of "center-bias" in human visual attention studies [6], [46]. It is mainly attributed to the fact that salient concepts mostly appear in the cental regions of images. Note that the model also attends to surrounding and lower regions at the following two timesteps, with the goal to find other concepts and their configurations at different locations.

Length of Semantic Order. For a given image, we need to manually set the number of timesteps T in the context-modulated attentional LSTM as the length of semantic order. Ideally, T should be equal to the number of semantic concepts appearing in the image. Therefore, the LSTM can separately attend to all the concepts within T steps to model their semantic order. To investigate what is the optimal length of semantic order, we gradually increase T from 1 to 5, and analyze the impact of different lengths on the performance on the Flick30k dataset in Table 4. From the table we observe that our model can achieve its best performance when the length of semantic order is 3. It indicates that it can capture all the concept-related regions by iteratively visiting the image for 3 times. Intuitively, when attending to an object, the LSTM can perceive both the object itself and its descriptive properties at one timestep. Considering that most images usually contain 2 major objects and a kind of relation (e.g., action), so the model additionally need two more timesteps.

Note that when T becomes larger than 3, the performance slightly drops. To investigate this, we show the predicted sequential attention maps when the length of semantic order is 6 in Fig. 8. We can see there mainly exist 3 classes of attention maps: 1) attended regions with individual concepts, 2) attended regions with multiple concepts, and 3) no particularly attended regions. The first class focuses on concepts of object and property, the second one focuses on relations among objects, and the last one focus on global scene. In fact, using 6 as the length of semantic order is unnecessary, since the third class of attention maps are redundant which attend to the less informative regions.

4.7 Effectiveness of Context-Modulated Attention

Attention Map Comparison. To qualitatively validate the effectiveness of using image context, we compare the generated attention maps without using image context in Fig. 5c. Without the aid of context, the model cannot produce accurate dynamical attention maps as those of using context. In particular, it cannot well attend to semantically meaningful concepts such as "cat" and "giraffe" with accurate outlines in the first and second images, respectively. In addition, it always finishes attending to concept-related regions within the first two timesteps, and does not focus on meaningful regions at the third timestep any more. These evidences show that the context modulation can be helpful for accurate concept discovery.

TABLE 4
The Performance of Different Lengths of the Semantic Order T

T	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
1	42.7	72.2	81.8	32.4	61.0	71.4	60.3
2	42.8	72.8	81.4	32.1	61.8	71.8	60.5
3	43.1	**73.4**	**83.1**	**32.9**	**63.9**	**74.3**	**61.8**
4	**44.0**	**73.4**	82.6	32.5	62.7	72.8	61.3
5	42.9	73.3	82.3	32.3	62.1	71.8	60.8

Gated Fusion Unit. The core of context modulation is the proposed gated fusion of semantic concepts and global scene, so it would be interesting to qualitatively analyze which images focus more on concepts or scene. Therefore, we plot the distribution of (sorted) weights on concepts for 1000 testing images of Flickr30k in Fig. 11. We can see that only a few images focus more on the scene than concepts, i.e., weight ≤ 0.5. We then find the visual content of these images are either too complex or ambiguous, and their predicted concepts are inaccurate. So these images tend to focus on the scene to extract more robust global information for cross-modal association.

Regularization Parameter. During the context-modulated attention, we add the regularization term to the structured and generation objectives, with the aim to force the model to pay equal attention to all the potential concept-related regions at different locations. We vary the values of regularization parameter μ from 0 to 1000, and compare the corresponding performance in Table 5. From the table, we can find that the performance improves when $\mu > 0$, which demonstrates the usefulness of paying attention to more diverse regions. In addition, when $\mu = 100$, the proposed model can achieve the largest performance improvement.

4.8 Evaluation of Joint Matching and Generation

Performance of Image Captioning. Since our model jointly performs sentence generation and matching, it can also be

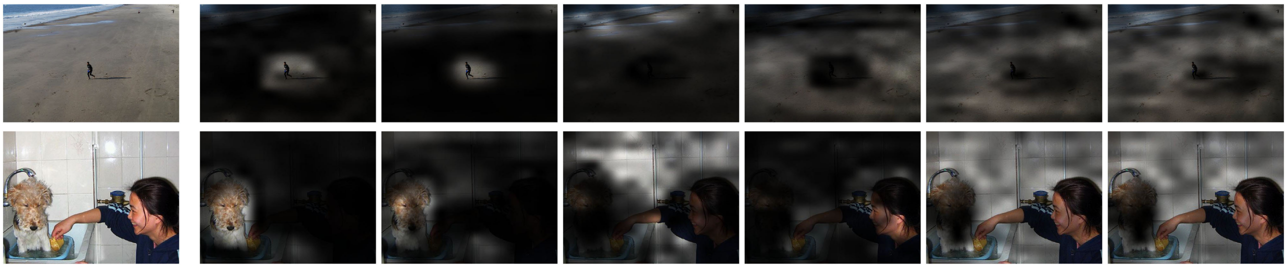


Fig. 8. Input images and attended image regions at six different timesteps (best viewed in colors).

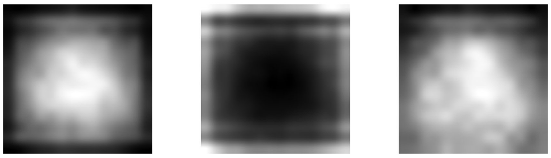

(a) 1-st timestep (b) 2-nd timestep (c) 3-rd timestep

Fig. 9. Averaged attention maps at three different timesteps.

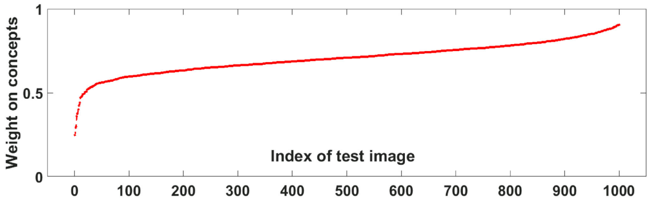


Fig. 11. The distribution of weights on concepts on Flickr30k.

applied to the task of image captioning. We present the results of image captioning on the Flickr30k dataset by our model in Table 6, and make comparisons with four closely related methods which share the same sentence generation module with ours. But they do not include the matching objective and use different modules of image representation, e.g., attention in SAT [57], global scene in ST [49], semantic attention in SA [59], and semantic concepts in CNP [55].

To demonstrate the effectiveness of 1) context-modulated attention for image representation improvement and 2) joint matching and generation for model learning, we preset the results of our two model variants: 1) a complete version with all the proposed modules (as "Ours"), and 2) an incomplete one without using the matching objective but only generation (as "Ours (w/o matching)"). Although without using the matching objective, Our model still achieves better performance than the compared models. It is mainly attributed to the use of context-modulated attention to learn more useful image representations. By jointly performing sentence matching and generation, the performance of our model can be further improved, which again verifies the effectiveness of our joint learning strategy.

Balancing Parameter. In addition, we test the balancing parameter λ between the structured and generation objectives in Equation (7). We vary it from 0 to 100 with a multiplier of 10, and present their corresponding results in Table 7. We can find that when $\lambda > 0$, we can always achieve better results. It shows that the use of sentence generation can help to enhance the learnt semantic order. When $\lambda = 1$, the model can achieve its best performance. It indicates that the generation objective plays an equally important role as the structured objective.

4.9 Error Analysis

To qualitatively illustrate the performance of our proposed model, we analyze the errors in its image annotation results as follows. We select several hard example images with complex content, and retrieve relevant sentences by 3 model variants: 1) "Ours (w/o concept and order)" uses only image scenes as image representations, 2) "Ours (w/o order)" improves image representations using predicted semantic concepts (in Fig. 6), and 3) "Ours" combines both concepts and their order. We show the retrieved top-5 relevant sentences by the 3 models in Fig. 10, and the predicted semantic concepts of query images in Fig. 6.

We can see that, without the aid of the predicted semantic concepts, our model cannot accurately capture the semantic concepts from complex image content. For example, the retrieved sentences contain some clearly wrong semantic concepts including *water* and *wine* for the first image, and lose important concepts such as *eating* and *basket* for the third image. After incorporating the predicted semantic concepts, the retrieved sentences have very similar meanings as the images, and are able to rank groundtruth sentences into top-5. But the top-1 sentences still do not involve partial image details or right semantic order, e.g., *bowl*, *sun* and *eating* for the three images, respectively. By further learning the semantic order with context-modulated attention and sentence generation, our model is able to retrieve the matched sentences with all the image details.

5 Conclusions and Future Work

In this work, we have proposed a semantic concepts and order learning framework for image and sentence matching.

Query	Retrieved top-5 relevant sentences		
	Ours (w/o concept or order)	Ours (w/o order)	Ours
	1. a dinner table with various plates of food and a glass of water on the table 2. a table top with some plates of food on it 3. a table set for three with food and wine 4. a dinner table with three plates of hamburgers 5. a table topped with plates of food and wine glasses	1. a meal is being displayed on a table **2. a table with bowls of grains and fruit and a hand with a plate** **3. a person holding a bowl of oats next to bowls of other condiments** 4. a table top with some plates of food on it 5. a table that has some food on it	**1. a person holding a bowl of oats next to bowls of other condiments** 2. a meal is being displayed on a table **3. a table with bowls of grains and fruit and a hand with a plate** 4. a table that has some food on it 5. table filled with a bunch of different types of food
	1. a man riding a skateboard up the side of a ramp 2. a man riding a skateboard up the side of a ramp 3. a man at a skate park with his foot on the side of the skateboard 4. a man on a skateboard performing a trick 5. a picture of a person on his skateboard	1. its a cloudy night for a ride on the motorcycle **2. a motorcyclist surveys the sunlit road into the horizon** **3. a close up of a person riding a motorcycle on a long empty road** 4. a photo taken from a car looking at a skateboarder on the side of the road **5. a photo taken from the back of a motorcycle cruising down the road**	**1. a motorcyclist surveys the sunlit road into the horizon** **2. a close up of a person riding a motorcycle on a long empty road** 3. a photo taken from a car looking at a skateboarder on the side of the road **4. a photo taken from the back of a motorcycle cruising down the road** 5. its a cloudy night for a ride on the motorcycle
	1. a couple of giraffes look around the ground 2. two giraffe standing near brick building 3. a pair of giraffes standing around in the enclosure 4. two giraffes roaming around an enclosed area 5. three giraffes standing in an enclosure overlooking Sydney Australia	1. a pair of giraffes standing around in their enclosure **2. a couple of giraffes eating hay from a trough** 3. two giraffes that are eating from a basket 4. two giraffes stand and eat food out of a basket 5. two giraffe standing next to each other near brick building	**1. a couple of giraffes eating hay from a trough** **2. a couple of giraffes eating out of a basket** 3. two giraffes stand and eat food out of a basket 4. a couple of giraffes reach for a basket 5. two giraffe standing next to each other near brick building

Fig. 10. Results of image annotation by 3 model variants. Groundtruth matched sentences are marked as red and bold, while some sentences sharing similar meanings as groundtruths are marked as underline (best viewed in colors).

TABLE 5
The Performance of Different Values of the Regularization Parameter μ

μ	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
0	40.4	70.4	80.4	29.6	61.9	72.7	59.2
1	41.4	71.6	81.9	30.7	62.5	73.1	60.2
10	42.5	72.7	82.8	31.5	62.9	73.8	61.0
100	**43.1**	**73.4**	**83.1**	**32.9**	**63.9**	**74.3**	**61.8**
1000	42.2	72.2	82.3	31.5	62.7	73.5	60.7

TABLE 6
Comparison Results of Image Captioning on the Flickr30k Dataset

Method	B-1	B-2	B-3	B-4	METEOR
ST [49]	0.66	0.42	0.28	0.18	-
SAT [57]	0.67	0.43	0.29	0.19	0.19
SA [59]	0.65	0.46	0.32	0.23	0.19
CNP [55]	0.73	0.55	0.40	0.28	-
Ours (w/o matching)	0.74	0.56	0.41	0.29	0.22
Ours	**0.77**	**0.58**	**0.43**	**0.31**	**0.23**

TABLE 7
The Performance of Different Values of the Balancing Parameter λ

λ	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
0	42.4	72.9	81.5	32.4	63.5	73.9	61.1
0.1	44.1	73.7	83.5	**32.8**	64.1	74.5	62.1
1	**44.2**	**74.1**	**83.6**	**32.8**	**64.3**	**74.9**	**62.3**
10	**44.2**	74.0	83.4	32.7	63.9	74.8	62.2
100	42.3	73.8	83.1	32.5	63.3	74.0	61.5

Our main contribution is improving the image representation by first predicting its semantic concepts and then organizing them in a correct semantic order. This is accomplished by multi-regional multi-label CNN, context-modulated attentional LSTM, and joint matching and generation learning, respectively. We have systematically studied the impact of these modules on the performance of image and sentence matching, and demonstrated the effectiveness of our model by achieving significant performance improvements. In the future, we will consider to generalize our model to the task of cross video-sentence learning, by learning semantic concepts and order for videos.

ACKNOWLEDGMENTS

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015), Beijing Natural Science Foundation (4162058), Capital Science and Technology Leading Talent Training Project (Z181100006318030), and Beijing Science and Technology Project (Z181100008918010). This work is also supported by Chinese Academy Sciences Artificial Intelligence Research (CAS-AIR).

REFERENCES

[1] T. D. Albright and G. R. Stoner, "Contextual influences on visual processing," *Annu. Rev. Neuroscience*, vol. 25, no. 1, pp. 339–379, 2002.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[3] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *Proc. Int. Conf. Learn. Representations*, 2015.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representations*, 2014.

[5] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 1171–1179.

[6] M. Bindemann, "Scene and screen center bias early eye movements in scene viewing," *Vis. Res.*, vol. 50, no. 23, pp. 2577–2587, 2010.

[7] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2422–2431.

[8] M. M. Chun and Y. Jiang, "Top-down attentional guidance based on implicit learning of visual covariation," *Psychological Sci.*, vol. 10, no. 4, pp. 360–365, 1999.

[9] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.

[10] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4601–4611.

[11] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improved visual-semantic embeddings," *British Mach. Vis. Conf. (BMVC)*, 2018.

[12] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al., "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1473–1482.

[13] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al., "Devise: A deep visual-semantic embedding model," in *Proc. Advances Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[14] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," arXiv:1312.4894, 2013.

[15] A. Graves, "Generating sequences with recurrent neural networks," arXiv:1308.0850, 2013.

[16] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1462–1471.

[17] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7181–7189.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] Y. Huang, W. Wang, and L. Wang, "Unconstrained multimodal multi-label learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1923–1935, no. 2015.

[21] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal LSTM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2310–2318.

[22] Y. Huang, Q. Wu, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6163–6171.

[23] A. Karpathy, A. Joulin, and F.-F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.

[24] A. Karpathy and F.-F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

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

[26] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4437–4446.

[27] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[29] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "RNN fisher vectors for action recognition and image annotation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 833–850.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[31] X. Lin and D. Parikh, "Leveraging visual question answering for image-caption ranking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 261–277.

[32] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4107–4116.

[33] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2623–2631.

[34] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," in *Proc. Int. Conf. Learn. Representations*, 2015.

[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Representations*, 2013.

[36] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 299–307.

[37] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends Cognitive Sci.*, vol. 11, no. 12, pp. 520–527, 2007.

[38] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6504–6512.

[39] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[40] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2641–2649.

[41] J. Pont-Tuset , P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, Jan. 2017.

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[43] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2014.

[45] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[46] P.-H. Tseng, R. Carmi, I. G. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. Vis.*, vol. 9, no. 7, pp. 4–4, 2009.

[47] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. Int. Conf. Learn. Representations*, 2016.

[48] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[49] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.

[50] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2285–2294.

[51] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5005–5013.

[52] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating human saccadic scanpaths on natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 441–448.

[53] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "CNN: Single-label to multi-label," arXiv:1406.5726, 2014.

[54] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3460–3469.

[55] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 203–212.

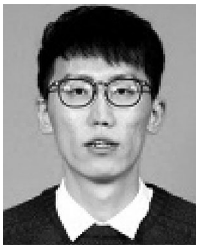
[56] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.

[57] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2048–2057.

[58] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3441–3450.

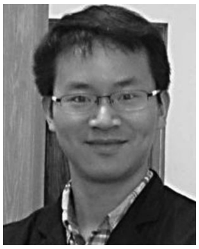
[59] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4651–4659.

[60] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014.



Yan Huang received the BSc degree from the University of Electronic Science and Technology of China (UESTC), in 2012, and the PhD degree from the University of Chinese Academy of Sciences (UCAS), in 2017. Since July 2017, He has joined the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) as an assistant professor. His research interests include machine learning and pattern recognition. He has published papers in the leading international journals and conferences such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Image Processing*, NIPS, CVPR, ICCV and ECCV.

Qi Wu received the bachelor of science degree in information and computing science from the China Jiliang University, and the PhD degree from the University of Bath, United Kingdom, in 2015. He is a lecturer at The University of Adelaides School of Computer Science. His research interests include in computer vision and natural language processing, and include image captioning, visual question answering and visual dialog. He publishes regularly in leading outlets such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE. Int. Conf. CVPR and IEEE Int. Conf. Computer Vision. He is currently a member of the Australian Institute for Machine Learning and the associate investigator of the Australia Centre for Robotic Vision.



Wei Wang received the BE degree from the Department of Automation from Wuhan University, in 2005, and the PhD degree from the School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences (GUCAS), in 2011. Since July 2011, he has joined NLPR as an assistant professor. His research interests focus on computer vision, pattern recognition and machine learning, particularly on the computational modeling of visual attention, deep learning and multimodal data analysis. He has published more than ten papers in the leading international conferences such as CVPR and ICCV.

Liang Wang received the BEng and MEng degrees from Anhui University, in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he was a research assistant at Imperial College London, United Kingdom, and Monash University, Australia, a research fellow with the University of Melbourne, Australia, and a lecturer with the University of Bath, United Kingdom, respectively. Currently, he is a full professor of the Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and the *IEEE Transactions on Image Processing*, and leading international conferences such as CVPR, ICCV, and ECCV. He is a fellow of the IEEE and the IAPR.