

Learning Semantic Concepts and Order for Image and Sentence Matching

Yan Huang^{1,3} Qi Wu⁴ Chunfeng Song^{1,3} Liang Wang^{1,2,3}

¹Center for Research on Intelligent Perception and Computing (CRIPAC),
 National Laboratory of Pattern Recognition (NLPR)

²Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
 Institute of Automation, Chinese Academy of Sciences (CASIA)

³University of Chinese Academy of Sciences (UCAS)

⁴School of Computer Science, The University of Adelaide

{yhuang, wangliang}@nlpr.ia.ac.cn qi.wu01@adelaide.edu.au

Abstract

Image and sentence matching has made great progress recently, but it remains challenging due to the large visual-semantic discrepancy. This mainly arises from that the representation of pixel-level image usually lacks of high-level semantic information as in its matched sentence. In this work, we propose a semantic-enhanced image and sentence matching model, which can improve the image representation by learning 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 semantic concepts, including objects, properties, actions, etc. Then, considering that different orders of semantic concepts lead to diverse semantic meanings, we use a context-gated sentence generation scheme for semantic order learning. It simultaneously uses the image global context containing concept relations as reference and the groundtruth semantic 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.

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,

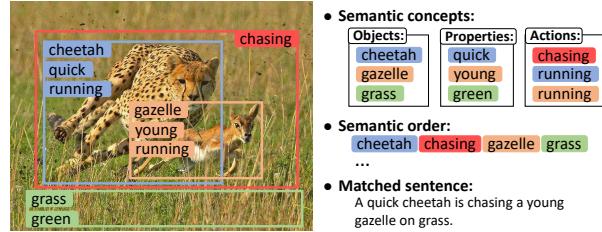


Figure 1. Illustration of the semantic concepts and order (best viewed in colors).

namely text-based image search.

Although much progress in this area has been achieved, it is still nontrivial to accurately measure the similarity between image and sentence, due to the existing huge visual-semantic discrepancy. Taking an image and its matched sentence in Figure 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 [12, 16, 23] jointly represent all the concepts by extracting a global CNN [30] feature vector, in which the concepts are tangled with each other. As a result, some primary foreground concepts 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 semantic concepts for the image, a possible way is to adaptively explore the attribute learning frameworks [6, 21, 35, 37]. But such a method 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 Figure 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 and matched sentence. But directly learning the correct semantic order from semantic concepts is very difficult, since there exist various incorrect orders that semantically make sense. We could resort to the image global context, since it already indicates the correct semantic order from the appearing spatial relations among semantic concepts, *e.g.*, the cheetah is on the left of the gazelle. But it is unclear how to suitably combine them with the semantic concepts, and make them directly comparable to the semantic order in the sentence.

Alternatively, we could generate a descriptive sentence from the image as its representation. However, the image-based sentence generation itself, namely image captioning, is also a very challenging problem. Even those state-of-the-art image captioning methods cannot always generate very realistic sentences that capture all image details. The image details are essential to the matching task, since the global image-sentence similarity is aggregated from local similarities in image details. Accordingly, these methods cannot achieve very high performance for image and sentence matching [3, 32].

In this work, to bridge the visual-semantic discrepancy between image and sentence, we propose a semantic-enhanced image and sentence matching model, which improves the image representation by learning semantic concepts and then organizing them in a correct semantic order. To learn the semantic concepts, we exploit a multi-regional multi-label CNN that can simultaneously predict multiple concepts in terms of objects, properties, actions, *etc*. The inputs of this CNN are multiple selectively extracted regions from the image, which can comprehensively capture all the concepts regardless of whether they are primary foreground ones. To organize the extracted semantic concepts in a correct semantic order, we first fuse them with the global context of the image in a gated manner. The context includes the spatial relations of all the semantic concepts, which can be used as the reference to facilitate the semantic order learning. Then we use the groundtruth semantic order in the matched sentence as the supervision, by forcing the fused image representation to generate the matched sentence.

After enhancing the image representation with both semantic concepts and order, we learn the sentence representation with a conventional LSTM [10]. Then the representations of image and sentence are matched with a structured objective, which is in conjunction with another objective of sentence generation for joint model learning. To demonstrate the effectiveness of the proposed model, we perform

several experiments of image annotation and retrieval on two publicly available datasets, and achieve the state-of-the-art results.

2. Related Work

2.1. Visual-semantic Embedding Based Methods

Frome *et al.* [7] propose the first visual-semantic embedding framework, in which ranking loss, CNN [17] and Skip-Gram [25] are used as the objective, image and word encoders, respectively. Under the similar framework, Kiros *et al.* [15] replace the Skip-Gram with LSTM [10] for sentence representation learning, Vendrov *et al.* [31] use a new objective that can preserve the order structure of visual-semantic hierarchy, and Wang *et al.* [34] additionally consider within-view constraints to learn structure-preserving representations.

Yan and Mikolajczyk [38] 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.* [16] use Fisher Vectors (FV) [27] to learn more discriminative representations for sentences, Lev *et al.* [18] alternatively use RNN to aggregate FV and further improve the performance, and Plummer *et al.* [28] explore the use of region-to-phrase correspondences. In contrast, our proposed model considers to bridge the visual-semantic discrepancy by learning semantic concepts and order.

2.2. Image Captioning Based Methods

Chen and Zitnick [2] use a multimodal auto-encoder for bidirectional mapping, and measure the similarity using the cross-modal likelihood and reconstruction error. Mao *et al.* [24] 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.* [3] design a long-term recurrent convolutional network for image captioning, which can be extended to image and sentence matching as well. Vinyals *et al.* [32] develop a neural image captioning generator and show the effectiveness on the image and sentence matching. 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.

3. Semantic-enhanced Image and Sentence Matching

In this section, we will detail our proposed semantic-enhanced image and sentence matching model from the following aspects: 1) sentence representation learning with a

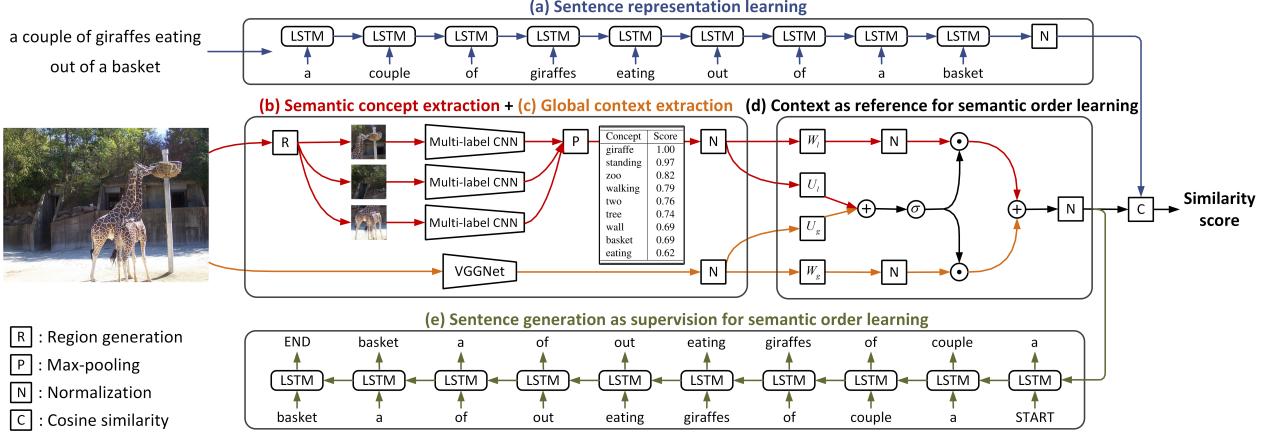


Figure 2. The proposed semantic-enhanced image and sentence matching model.

conventional LSTM, 2) semantic concept extraction with a multi-regional multi-label CNN, 3) semantic order learning with a context-gated sentence generation scheme, and 4) model learning with joint image and sentence matching and sentence generation.

3.1. Sentence Representation Learning

For a sentence, its included nouns, verbs and adjectives directly correspond to the visual semantic concepts of object, property and action, respectively, which are already given. 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 [15, 31]. The LSTM has multiple components for information memorizing and forgetting, which can well suit the complex properties of semantic concepts and order. As shown in Figure 2 (a), we sequentially feed all the words of the sentence into the LSTM at different timesteps, and then regard the hidden state at the last timestep as the desired sentence representation $\mathbf{s} \in \mathbb{R}^H$.

3.2. Image Semantic Concept Extraction

For images, their semantic concepts refer to various objects, properties, actions, *etc.* The existing datasets do not provide these information 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 [6, 37]. 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 the concept vocabulary is very large, we ignore those words that have very low use frequencies. In addition, we unify the different tenses of verbs, and the singular and plural forms of nouns 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 [8, 11, 13, 33, 35, 36, 37], which mostly learn various CNN-based models as nonlinear mappings from images to the desired multiple labels. Similar to [35, 37], we simply use the VGGNet [30] pre-trained on the ImageNet dataset [29] 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. 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 $\hat{\mathbf{y}}_i \in [0, 1]^K$, then the model can be learned by optimizing the following objective:

$$L_{cnn} = \sum_{c=1}^K \log(1 + e^{(-\mathbf{y}_i, \hat{\mathbf{y}}_i)_c}) \quad (1)$$

During testing, considering that the semantic concepts usually appear in image local regions and vary in size, we perform the concept prediction in a regional way. Given a testing image, we first selectively extract r image regions in a similar way as [35], and then resize them to square shapes. As shown in Figure 2 (b), by separately feeding these regions into the learned multi-label CNN, we can obtain a set of predicted confidence score vectors. Note that the model parameters are shared among all the regions. 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.

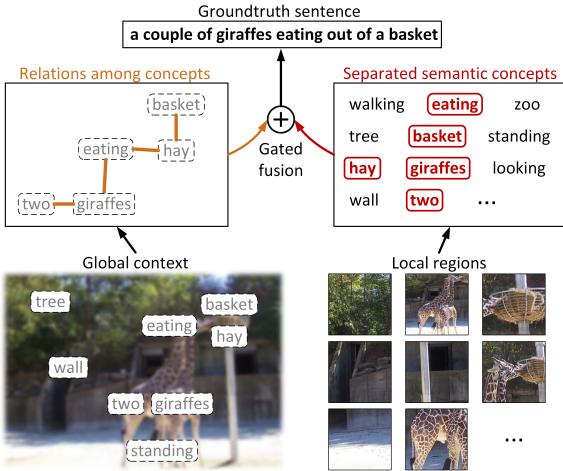


Figure 3. Illustration of using the global context as reference for semantic order learning (best viewed in colors).

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 Figure 2 (b), if we organize the extracted semantic concepts: *giraffes*, *eating* and *basket* as: *a basket is eating two giraffes*, then its meaning is very different from the image content. To learn the semantic order, we propose a context-gated sentence generation scheme that uses the image global context as reference and the sentence generation as supervision.

3.3.1 Global Context as Reference

It is not easy to learn the semantic order directly from separated semantic concepts, since the semantic order involves not only the hypernym relations between concepts, but also the textual entailment among phrases in high levels of semantic hierarchy [31]. To deal with this, we propose to use the image global context as auxiliary reference for semantic order learning. As illustrated in Figure 3, the global context can not only describe all the semantic concepts in a coarse level, but also indicate their spatial relations with each other, e.g., two giraffe are standing in the left while the basket is in the top left corner. When organizing the separated semantic concepts, our model can refer to the global context to find their relations and then combine them to facilitate the prediction of semantic order. 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 context, as shown in Figure 2 (c).

To model such a reference procedure, a simple way is to

sum the global context with semantic concepts together. But considering that the content of different images can be diverse, thus the relative importance of semantic concepts and context is not equivalent in most cases. For those images with complex content, their global context might be a bit of ambiguous, so the semantic concepts are more discriminative. To handle this, we design a gated fusion unit that can selectively balance the relative importance of semantic concepts and context. The unit acts as a gate that controls how much information of the semantic concepts and context contributes to their fused representation. As illustrated in Figure 2 (d), after obtaining the normalized context vector $\mathbf{x} \in \mathbb{R}^I$ and concept score vector $\mathbf{p} \in \mathbb{R}^K$, their fusion by the gated fusion unit can be formulated as:

$$\begin{aligned}\hat{\mathbf{p}} &= \|W_l \mathbf{p}\|_2, \quad \hat{\mathbf{x}} = \|W_g \mathbf{x}\|_2, \quad \mathbf{t} = \sigma(U_l \mathbf{p} + U_g \mathbf{x}) \\ \mathbf{v} &= \mathbf{t} \odot \hat{\mathbf{p}} + (1 - \mathbf{t}) \odot \hat{\mathbf{x}}\end{aligned}\quad (2)$$

where $\|\cdot\|_2$ denotes the l_2 -normalization, and $\mathbf{v} \in \mathbb{R}^H$ is the fused representation of semantic concepts and global context. 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 \mathbf{v} becomes an element-wise weighted sum of \mathbf{p} and \mathbf{x} .

3.3.2 Sentence Generation as Supervision

To learn the semantic order based on the fused representation, a straightforward approach is to directly generate a sentence from it, similar to image captioning [37]. However, such an approach is infeasible resulting from the following problem. Although the current image captioning methods can generate semantically meaningful sentences, the accuracy of their generated sentences on capturing image details is not very high. And even a little error in the sentences can be amplified and further affect the measurement of similarity, since the generated sentences are highly semantic and the similarity is computed in a fine-grained level. Accordingly, even the state-of-the-art image captioning models [3, 24, 32] cannot perform very well on the image and sentence matching task. We also implement a similar model (as “ctx + sen”) in Section 4.3, but find it only achieves inferior results.

In fact, it is unnecessary for the image and sentence matching task to generate a grammatically-complete sentence. We can alternatively regard the fused context and concepts as the image representation, and supervise it using the groundtruth semantic order in the matched sentence during the sentence generation. As shown in Figure 2 (e), 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 $\{\mathbf{w}_j | \mathbf{w}_j \in \{0, 1\}^G\}_{j=1, \dots, J}$, 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:

$$\begin{aligned}\mathbf{i}_t &= \sigma(W_{\mathbf{wi}}(F\mathbf{w}_t) + W_{\mathbf{hi}}\mathbf{h}_{t-1} + \mathbf{b}_i), \\ \mathbf{f}_t &= \sigma(W_{\mathbf{wf}}(F\mathbf{w}_t) + W_{\mathbf{hf}}\mathbf{h}_{t-1} + \mathbf{b}_f), \\ \mathbf{o}_t &= \sigma(W_{\mathbf{wo}}(F\mathbf{w}_t) + W_{\mathbf{ho}}\mathbf{h}_{t-1} + \mathbf{b}_o), \\ \hat{\mathbf{c}}_t &= \tanh(W_{\mathbf{wc}}(F\mathbf{w}_t) + W_{\mathbf{hc}}\mathbf{h}_{t-1} + \mathbf{b}_c), \\ \mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t, \quad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \\ \mathbf{q}_t &= \text{softmax}(F^T \mathbf{h}_t + \mathbf{b}_p), \quad e = \arg \max(\mathbf{w}_t), \\ P(\mathbf{w}_t | \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \dots, \mathbf{w}_0, \mathbf{x}, \mathbf{p}) &= \mathbf{q}_{t,e}\end{aligned}\quad (3)$$

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 context \mathbf{x} at the initial timestep.

3.4. Joint Matching and Generation

During the model learning, to jointly perform image and sentence matching and sentence generation, we need to minimize the following combined objectives:

$$L = L_{\text{mat}} + \lambda \times L_{\text{gen}} \quad (4)$$

where λ is a tuning parameter for balancing.

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 i -th image and i -th sentence, s_{ik} is the score of mismatched i -th image and k -th 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 context \mathbf{x} :

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

where the detailed formulation of probability P is shown in Equation 3. Note that we use the predicted semantic concepts rather than groundtruth ones in our experiments.

All modules of our model excepting for the multi-regional multi-label CNN can constitute a whole deep network, which can be jointly trained in an end-to-end manner from raw image and sentence to their similarity score. It should be noted that we do not need to generate the sentence during testing. We only have to compute the image representation \mathbf{v} from \mathbf{x} and \mathbf{p} , and then compare it with the

sentence representation \mathbf{s} to obtain their cosine 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 [39] 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 [15], which contain 28000, 1000 and 1000 images, respectively. 2) MSCOCO [19] 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 [15], 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.

For images, the dimension of global context is $I=4096$ for VGGNet [30] or $I=1000$ for ResNet [9]. We perform 10-cropping [16] from the images and then separately feed the cropped regions into the network. The final global context is averaged over 10 regions. For sentences, the dimension of embedded word is $D=300$. We set the max length for all the sentences as 50, *i.e.*, the number of words $J=50$, and use zero-padding when a sentence is not long enough. Other parameters are empirically set as follows: $H=1024$, $K=256$, $\lambda=1$, $r=50$ and $m=0.2$.

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) “1-crop” and “10-crop” refer to cropping 1 or 10 regions from images, respectively, when extracting the global context. 2) “concept” and “context” denote using semantic concepts and global context, respectively. 3) “sum” and “gate” are two different ways that combine semantic concepts and context via feature summation and gated fusion unit, respectively. 4) “sentence”, “generation” and “sampling” are three different ways to learn the semantic order, in which “sentence” uses the state-of-the-art image captioning method [32] to generate sentences

Table 1. The experimental settings of ablation models.

	1-crop	10-crop	context	concept	sum	gate	sentence	generation	sampling	shared	non-shared
ctx (1-crop)	✓		✓								
ctx		✓	✓								
ctx + sen		✓	✓				✓				
ctx + gen (S)		✓	✓					✓	✓		
ctx + gen (E)		✓	✓					✓		✓	
ctx + gen		✓	✓					✓			✓
cnp				✓							
cnp + gen				✓				✓			
cnp + ctx (C)		✓	✓	✓	✓						
cnp + ctx		✓	✓	✓	✓	✓					
cnp + ctx + gen		✓	✓	✓	✓	✓		✓			✓

Table 2. 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	
ctx (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
ctx	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
ctx + sen	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
ctx + gen (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
ctx + gen (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
ctx + gen	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
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 + gen	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
cnp + ctx (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 + ctx	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
cnp + ctx + gen	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

from images and then regard the sentences as the image representations, “generation” uses the sentence generation as supervision as described in Section 3.3.2, and “sampling” additionally uses the scheduled sampling [1]. 5) “share” and “non-shared” indicate whether the parameters of two word embedding matrices for sentence representation learning and sentence generation are shared or not.

4.3. Evaluation of Ablation Models

The results of the ablation models on the Flickr30k and MSCOCO datasets are shown in Table 2, from which we can obtain the following conclusions. 1) Cropping 10 image regions (as “ctx”) can achieve much robust global context features than cropping only 1 region (as “ctx (1-crop)”). 2) Directly using the pre-generated sentences as image representations (as “ctx + sen”) cannot improve the performance, since the generated sentences might not accurately include the image details. 3) Using the sentence generation as supervision for semantic order learning (as “ctx + gen”) is very effective. But additionally performing the scheduled sampling (as “ctx + gen (S)”) cannot further improve the performance. It is probably because the groundtruth semantic order is degenerated during sampling, accordingly the model cannot learn it well. 4) Using a shared word embedding matrix (as “ctx + gen (E)”) cannot improve the performance, which might result from that learning a unified matrix for two tasks is difficult. 5) Only using the se-

mantic concepts (as “cnp”) can already achieve good performance, especially when the training data are sufficient on the MSCOCO dataset. 6) Simply summing the concept and context (as “cnp + ctx (C)”) can further improve the result, because the context contains the spatial relations of concepts which are very useful. 7) Using the proposed gated fusion unit (as “cnp + ctx”) performs better, due to the effective importance balancing scheme. 8) The best performance is achieved by the “cnp + ctx + gen”, which combines the 10-cropped extracted context with semantic concepts via the gated fusion unit, and exploits the sentence generation for semantic order learning. Without using either semantic concepts (as “ctx + gen”) or context (as “cnp + gen”), the performance drops heavily. In the follow experiments, we regard the “cnp + ctx + gen” as our default model.

In addition, we test the balancing parameter λ in Equation 4, by varying it from 0 to 100. The corresponding results are presented in Table 3, we can find that when $\lambda=1$, the model can achieve its best performance. It indicates that the generation objective plays an equally important role as the matching objective.

4.4. Comparison with State-of-the-art Methods

We compare our proposed model with several recent state-of-the-art models on the Flickr30k and MSCOCO datasets in Table 4. The methods marked by “(Res)” use the 152-layer ResNet [9] for context extraction, while the

Table 3. Comparison results of image annotation and retrieval on the MSCOCO (1000 testing) dataset.

λ	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		R@1	R@5	R@10	
0	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				
0.01	43.1	72.8	83.5	32.8	63.2	73.6	61.5	66.3	91.2	96.5	55.4	86.5	93.7	81.6				
1	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				
100	42.3	73.8	83.1	32.5	63.3	74.0	61.5	65.0	90.5	96.1	54.9	86.3	93.7	81.1				

Table 4. Comparison results of image annotation and retrieval 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		R@1	R@5	R@10	
m-RNN [24]	35.4	63.8	73.7	22.8	50.7	63.1	51.6	41.0	73.0	83.5	29.0	42.2	77.0	57.6				
FV [16]	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 [14]	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 [15]	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 [23]	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 [18]	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 [31]	-	-	-	-	-	-	-	46.7	78.6	88.9	37.9	73.7	85.9	68.6				
VQA [20]	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 [28]	37.4	63.1	74.3	26.0	56.0	69.3	54.3	-	-	-	-	-	-	-	-	-	-	
DSPE [34]	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 [12]	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 [4]	49.8	67.5	-	36.0	55.6	-	-	55.8	75.2	-	39.7	63.3	-	-	-	-	-	
DAN [26]	41.4	73.5	82.5	31.8	61.7	72.5	60.6	-	-	-	-	-	-	-	-	-	-	
VSE++ [5]	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				
Ours	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				
RRF (Res) [22]	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) [26]	55.0	81.8	89.0	39.4	69.2	79.1	68.9	-	-	-	-	-	-	-				
VSE++ (Res) [5]	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				
Ours (Res)	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				

rest ones use the default 19-layer VGGNet [30].

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.

The above experiments on the MSCOCO dataset follow the first protocol [14], 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 5. From the table we can observe that the overall results by all the methods are lower than the first protocol. It probably results from that the target set is much larger so there exist more distracters for a given query. Among all the model-

Table 5. 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 [14]	11.8	32.5	45.4	8.9	24.9	36.3	26.6
FV[16]	17.3	39.0	50.2	10.8	28.3	40.1	31.0
VQA [20]	23.5	50.7	63.6	16.7	40.5	53.8	41.5
OEM [31]	23.3	50.5	65.0	18.0	43.6	57.6	43.0
VSE++ [5]	32.9	61.6	74.7	24.1	52.0	66.2	51.9
Ours	40.2	70.1	81.3	31.3	61.5	73.9	59.7
VSE++ (Res) [5]	41.3	69.2	81.2	30.3	59.1	72.4	58.9
Ours (Res)	42.8	72.3	83.0	33.1	62.9	75.5	61.6

s, the proposed model still achieves the best performance, which again demonstrates its effectiveness. Note that our model has much larger improvements using VGGNet than ResNet, which results from that “Ours (Res)” only uses the ResNet for extracting global context but not semantic concepts.

4.5. Analysis of Image Annotation Results

To qualitatively validate the effectiveness of our proposed model, we analyze its image annotation results as follows. We select several representative images with complex content, and retrieve relevant sentences by 3 ablation models: “ctx”, “cnp + ctx” and “cnp + ctx + gen”. We show

Query	Retrieved top-3 relevant sentences		
	ctx	cnp + ctx	cnp + ctx + gen
① 	1. a dinner table with various plates of food and a glass of water on the table <u>2. a table top with some plates of food on it</u> 3. a table set for three with food and wine	<u>1. a meal is being displayed on a table</u> 2. a table with bowls of grains and fruit and a hand with a plate <u>3. a table top with some plates of food on it</u>	1. a person holding a bowl of oats next to bowls of other condiments <u>2. a meal is being displayed on a table</u> <u>3. a table with bowls of grains and fruit and a hand with a plate</u>
② 	1. a man on a skateboard performing a trick 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	1. its a cloudy night for a ride on the motorcycle 2. a motorcyclist surveys the sunlit road into the horizon <u>3. a close up of a person riding a motorcycle on a long empty road</u>	1. a motorcyclist surveys the sunlit road into the horizon <u>2. a close up of a person riding a motorcycle on a long empty road</u> 3. a photo taken from a car looking at a skateboarder on the side of the road
③ 	1. a couple of giraffes look around the ground in the zoo 2. two giraffe standing near brick building 3. a pair of giraffes standing around in their enclosure	1. a pair of giraffes standing around in their enclosure 2. a couple of giraffes eating hay from a trough <u>3. two giraffes that are eating from a basket</u>	1. a couple of giraffes eating hay from a trough <u>2. a couple of giraffes eating out of a basket</u> <u>3. two giraffes stand and eat food out of a basket</u>

Figure 4. Results of image annotation by 3 ablation models. 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).

the retrieved top-3 relevant sentences by the 3 models in Figure 4, and the predicted top-10 semantic concepts with confidence scores in Figure 5.

From Figure 5, we can see that our multi-regional multi-label CNN can accurately predict the 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 the reason that this image content is complicated and the smooth country road looks like some skating scenes.

As shown in Figure 4, without the aid of the predicted semantic concepts, “ctx” 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 by “cnp + ctx” have very similar meanings as the images, and are able to rank groundtruth sentences into top-3. But the top-1 sentences still do not involve partial image details, e.g., *bowl*, *sun* and *eating* for the three images, respectively. By further learning the semantic order with sentence generation, the “cnp + ctx + gen” is able to associate all the related concepts and retrieve the matched sentences with all the image details.

5. Conclusions and Future Work

In this work, we have proposed a semantic-enhanced image and sentence matching model. Our main contribution is improving the image representation by learning semantic concepts and then organizing them in a correct semantic order. This is accomplished by a series of model components in terms of multi-regional multi-label CNN, gated fu-

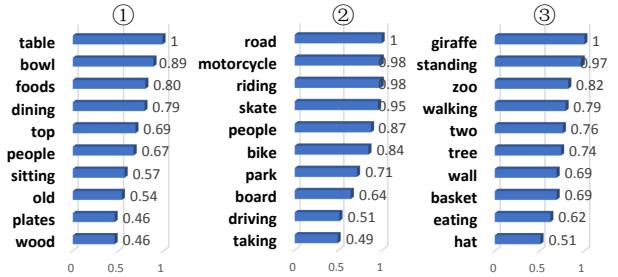


Figure 5. Predicted top-10 semantic concepts with their confidence scores from the 3 images.

sion unit, and joint matching and generation learning. We have systematically studied the impact of these components on the image and sentence matching, and demonstrated the effectiveness of our model by achieving significant performance improvements.

In the future, we will replace the used VGGNet with ResNet in the multi-regional multi-label CNN to predict the semantic concepts more accurately. Our model can perform image and sentence matching and sentence generation, so we would like to extend it for the image captioning task.

Acknowledgements

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), and Capital Science and Technology Leading Talent Training Project (Z181100006318030). This work is also supported by grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer.

References

- [1] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, 2015.
- [2] X. Chen and C. Lawrence Zitnick. Mind’s eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015.
- [3] 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 *CVPR*, 2015.
- [4] A. Eisenschatz and L. Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017.
- [5] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. Vse++: Improved visual-semantic embeddings. *arXiv*, 2017.
- [6] 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 *CVPR*, 2015.
- [7] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.
- [8] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv*, 2013.
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.
- [11] Y. Huang, W. Wang, and L. Wang. Unconstrained multimodal multi-label learning. *IEEE TMM*, 2015.
- [12] Y. Huang, W. Wang, and L. Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, 2017.
- [13] Y. Huang, W. Wang, L. Wang, and T. Tan. Multi-task deep neural network for multi-label learning. In *ICIP*, 2013.
- [14] A. Karpathy and F.-F. Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- [15] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015.
- [16] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015.
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [18] G. Lev, G. Sadeh, B. Klein, and L. Wolf. Rnn fisher vectors for action recognition and image annotation. In *ECCV*, 2016.
- [19] 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 *ECCV*. 2014.
- [20] X. Lin and D. Parikh. Leveraging visual question answering for image-caption ranking. In *ECCV*, 2016.
- [21] J. Liu, L. Wang, M.-H. Yang, et al. Referring expression generation and comprehension via attributes. In *ICCV*, 2017.
- [22] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew. Learning a recurrent residual fusion network for multimodal matching. In *ICCV*, 2017.
- [23] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015.
- [24] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. In *ICLR*, 2015.
- [25] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.
- [26] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, 2017.
- [27] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
- [28] 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 *ICCV*, 2015.
- [29] 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. *IJCV*, 2015.
- [30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014.
- [31] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.
- [32] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE TPAMI*, 2017.
- [33] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, 2016.
- [34] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.
- [35] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Cnn: Single-label to multi-label. *arXiv*, 2014.
- [36] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *CVPR*, 2015.
- [37] 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 *CVPR*, 2016.
- [38] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015.
- [39] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014.