This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2811621, IEEE Transactions on Multimedia

Bundled Object Context for Referring Expressions

Xiangyang Li and Shuqiang Jiang, *Senior Member, IEEE*

***Abstract*—Referring expressions are natural language descriptions of objects within a given scene. Context is of crucial importance for a referring expression as the description not only depicts the properties of the object, but also involves the relationships of the referred object with other ones. Most of previous work uses either the whole image or one particular contextual object as the context. However, the context of these approaches is holistic and insufficient, as a referring expression often describes relationships of multiple objects in an image. To leverage rich context information from all objects in an image, in this work, we propose a novel scheme which is composed of a visual context Long Short-Term Memory (LSTM) module and a sentence LSTM module to model bundled object context for referring expressions. All contextual objects are arranged with their spatial locations and progressively fed into the visual context LSTM module to acquire and aggregate the context features. And then the concatenation of the learned context features and the features of the referred object are put into the sentence LSTM module to learn the probability of a referring expression. The feedback connections and internal gating mechanism of the LSTM cells enable our model to selectively propagate relevant contextual information through the whole network. Experiments on three benchmark datasets show our methods can achieve promising results compared to state-of-the-art methods. Moreover, visualization of the internal states of the visual context LSTM cells also shows that our method can automatically select the pertinent context objects.**

***Index Terms*—Bundled object context, referring expression, LSTM, vision-language.**

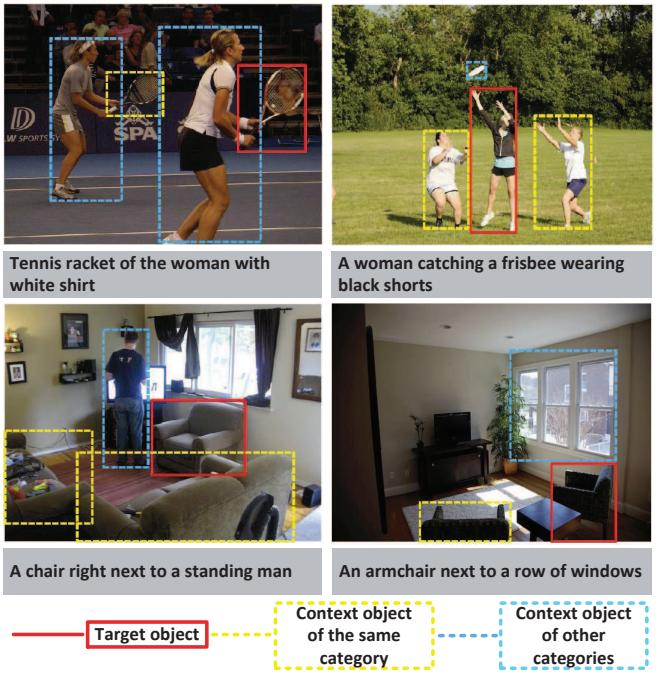


Fig. 1. Complex context in referring expression comprehension and generation (Best viewed in color).

I. Introduction

VISION and language are intrinsically two different ways to represent and exchange information in our daily life. Recently, how to bridge these two domains has been paid many research attentions in the area of multimedia [1], [2], [3], [4], [5] , computer vision [6], [7], [8], [9] and natural language processing [10], [11]. The majority of previous work which has focused on linguistic description of images can be categorized into two types. The first one is image captioning by describing the entire image with one sentence [7], [12], [13], [14], [15]. However, it is usually insufficient to describe an image just with one sentence. Simultaneously, with many valid ways to describe a given image, it is hard to evaluate one description is better than the others. To address these deficiencies, the second type describes an image at the region level, such as dense captioning [16], grounding textual phrases [17], referring expression comprehension and generation [18], [19], [20], [21] and so on. While dense captioning is to describe all salient objects in an image, and grounding textual phrases is to align image regions and textual phrases in an image-sentence pair, *referring expressions* are natural language expressions which describe an object or region in an image, for the purpose of pinpointing it uniquely to a listener.

As humans often refer to objects in the physical world when they have a pragmatic interaction with others, being able to generate and comprehend referring expression has a wide range of practical applications. For example, in human-robot interaction, a robot dialogue system needs to understand expressions referring to objects in its surrounding environments. The ability to understand referring expressions makes the system can handle natural language commands such as passing me “the air conditioning remote control on the shelf”, picking up “the green bolt on the table”, *etc.* Another example is that understanding referring expressions enables hand-free mechanism to select objects of interest in an image. For example, one can tell an image editing software to refine the properties of one entity in an image by the instruction of “the plant on the right side of the TV” or “the umbrella held by a woman wearing a blue jacket”, replacing the traditional way by indicating it with the mouse.

This work was supported in part by the National Natural Science Foundation of China under Grant 61532018, in part by the Beijing Municipal Commission of Science and Technology under Grant D161100001816001, in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals.

X. Li and S. Jiang are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, are also with University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: xiangyang.li@vipl.ict.ac.cn; sqjiang@ict.ac.cn).

Within the realm of referring expressions, in this paper, we consider two related tasks of *referring expression comprehension* and *referring expression generation*, mimicking the listener and speaker roles. (1) The comprehension task mimics the role of a listener to localize an object in an image given a referring expression. (2) The inverse task is the generation task mimicking the role of a speaker to generate a discriminative referring expression for an object in an image.

Referring expressions usually not only describe the properties such as color, texture, size and location of the referred object itself, but also depict its relationships with other objects in the given image. Moreover, in contrast to generic context-agnostic image captions, referring expressions are context-aware. For instance, in the top left of Fig. 1, a literal description "A tennis racket held by a woman" conveys abundant semantic information to describe the target tennis racket (bounded with red solid rectangle), but it would be inadequate to disambiguate it from another one within the image. When generating an expression for a target object, we inherently emphasize the referred object while keeping other context objects within the image in mind.

Above all, to generate a satisfactory expression, context is of crucial importance. It is an essential part of the input information to provide a broader understanding of the image and deliver visual cues to decide what to emphasize and what to ignore. Some previous work has studied the problem of modeling context for referring expressions. These methods either use the whole image [22] or one particular context object [23] as the context. However, this kind of context is holistic and insufficient, as there are usually multiple objects in an image and a referring expression often involves multiplex relationships between them. We human beings always take many context objects into consideration when generating an expression for a target object. For example, as shown in Fig. 1, when generating the expression "Tennis racket of the woman with white shirt", besides considering how to describe the target tennis racket (bounded with red solid rectangle), we not only take the attributes of the other tennis racket (bounded with yellow dash rectangle) into account, but also infer how to distinguish two instances of woman (bounded with cyan solid rectangle) in the image. Both of these related context objects provide information to produce the appropriate results.

To generate an unambiguous expression to describe an object, speakers perform an exhaustive scan of the objects in the image [24]. In the same spirit, we exploit multiple objects context for referring expressions by leveraging all of the objects in the image. Unlike the work of Yu *et al.* [25], which just focus on visual comparisons among the other objects of the same category of the target object, we use objects of both the same category and different categories with the target object. By considering all of these objects, the learned context reveals more details of the image and provides larger visual cues and abundant information to describe the target object.

In this paper, we propose a novel framework to model bundled object context for referring expressions, as shown in Fig. 2. All contextual objects are bundled together with an Long Short-Term Memory (LSTM) network to get the context features. To represent the objects in a sequential way, we arrange them in an order of their locations in the image. The LSTM cell progressively takes each of the objects as inputs and decides whether to retain the information from the current input or discard it based on the information it captured from its previous states and current input. Then the learned bundled object context features and the features of the referred object are put into a second LSTM to learn the probability of a referring expression. The feedback connections and internal gating mechanism of the LSTM cells enable our model to selectively propagate relevant contextual information. Experiments on three benchmark datasets show our methods can achieve promising results compared to state-of-the-art methods. Visualization of the internal states of the LSTM cells also show that our method can automatically select the pertinent context objects.

In summary, the main contributions of our paper are as follows:

- We firstly use an LSTM network to learn bundled object context for referring expressions from all context objects in the image. The learned context contains information from objects of both the same category and different categories with the target object.
- We visualize the internal states of the LSTM cells. The results show that our model can automatically select the relevant context objects to capture discriminative and informative context features.
- We quantitatively and qualitatively validate the effectiveness of our model on three benchmark datasets, and achieve promising results compared to state-of-the-art methods.

In the following of our paper, we give a brief overview of related work in Section II. We then describe our methods for referring expression comprehension and generation in Section III and present the quantitative and qualitative results in Section IV. At last, we give our conclusion in Section V.

II. Related Work

In this section, we briefly review the related work on image captioning and referring expressions. Moreover, we also review a series of methods using recurrent neural networks to capture rich contextual information.

A. Image Captioning

The task of image captioning takes an entire image as the input and outputs a natural language sentence that describes its content. Farhadi *et al.* [26] use the method based on image retrieval. That is, to create a description for a novel image, they first search captioned images which are similar to the novel image in a database, and then simply transfer the descriptions of the retrieved images to the novel one. Kulkarni *et al.* [6] use the method generating descriptions based on a set of fixed sentence templates. They first detect elements such as attributes, objects and actions in the image, and then fill the template with them to generate sentences. Recent popular approaches are based on neural networks. Karpathy *et al.* [7] and Mao *et al.* [27] use a deep Convolutional

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2811621, IEEE Transactions on Multimedia

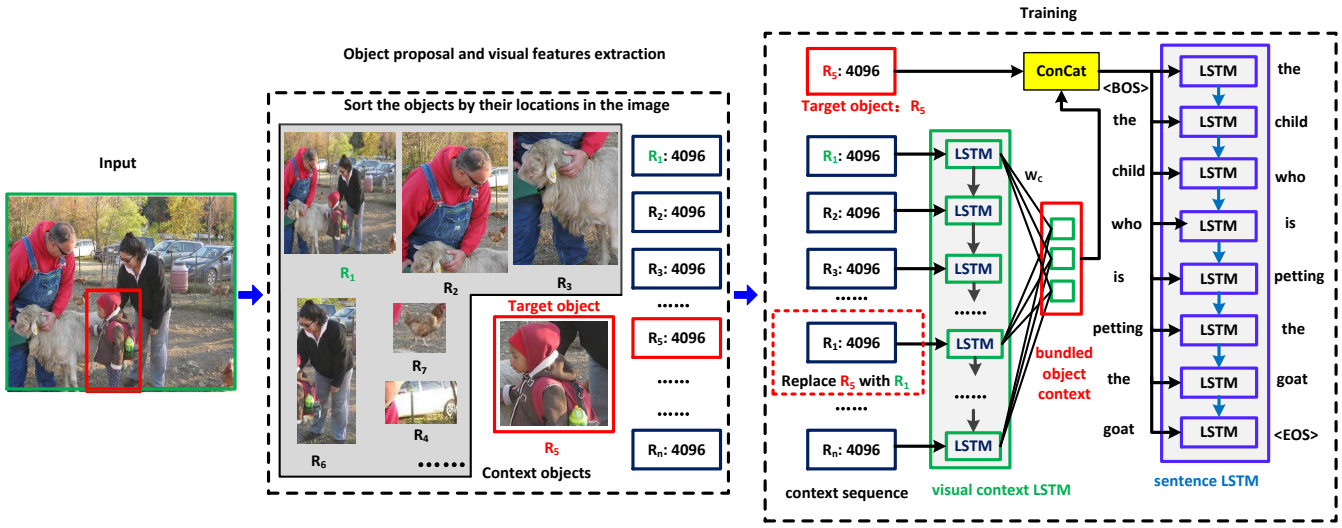


Fig. 2. The architecture diagram of our method. Given an image, we first detect a fixed number of objects in the image. These objects are sorted by their locations in the image with the order of from left to right and top to down (The entire image is also considered as a special object, denoted as R_1). The visual context LSTM initially processes the context sequence which is composed of all the context objects to produce their hidden representations. All the hidden representations are combined with a fully connected layer to obtain the bundled object context. The sentence LSTM receives the features of the target object as well as the context features to generate the description.

Neural Network (CNN) to extract visual features, and then put the features into a Recurrent Neural Network (RNN) as the initial start word to generate image descriptions. Wu *et al.* [28] propose a method that explicitly incorporates high-level concepts into the successful CNN-RNN based approach, and show that it achieves a significant improvement. Xu *et al.* [13], You *et al.* [29] and Liu *et al.* [30] incorporate attention mechanism to LSTM networks to generate sentences. Jin *et al.* [12] use attention based on image regions to generate global descriptions for images. While attention is sequentially used to select visual or semantic elements when generating one token of a description, the proposed method in our paper aims to model useful context when generating the whole description. Even though much progress has been achieved in image captioning, one challenge is that the description that a system should generate for an whole image is task dependent. A further step beyond image captioning is to locate or describe image regions. Therefore, there have been many tasks focusing on image regions or objects such as dense captioning [16], grounding textual phrases [17]. In this paper, we focus on referring expression generation and comprehension which also involve specific objects within an image.

B. Referring Expressions

Referring expressions have attracted research interest in multimedia and related areas. Comprehension and generation are two tasks associated with referring expression. The comprehension task requires a system to select the object described by a given expression. To address this issue, Mao *et al.* [22] and Hu *et al.* [31] first learn a generative model with paired object and sentence and then look for object maximizing the probability of the given expression. Wang *et al.* [32] and Plummer *et al.* [33] learn embedding models to minimize the distance between paired object and text. The generation task is to generate an expression for a specified object within an image. Many approaches have been proposed for referring expression generation [34], [19], [24], [35], [36]. FitzGerald *et al.* [20] learn a probabilistic model for referring expressions which identify sets of visual objects, and treat the task as a density estimation problem. Kazemzadeh *et al.* [18] use an optimization based method for referring expressions generation. Different from these approaches which use template-based methods to generate expressions with fixed grammar pattern, Mao *et al.* [22] and Hu *et al.* [31] begin to use LSTM based networks to generate referring expressions. Luo *et al.* [35] propose to use learned comprehension models to guide generating better referring expressions. As generating referring expressions is concerned with how we produce a description of an object that enables the listener to identify that object in a given image, the context is very important. Unlike most of the works only using the whole image as the context, Nagaraja *et al.* [23] explicitly use a multiple-instance learning method to learn a supporting context object for each target object to comprehend referring expressions. Yu *et al.* [25] add visual comparisons which encode the visual difference among the other objects of the same category of the target object to the context features. In this work, we use an LSTM to learn bundled object context. Our context is informative because it is composed of objects of both the same category and different categories with the target object in the given image.

C. Modeling Context with RNNs

The recurrent connections with its previous states allow Recurrent Neural Network (RNN) to have the ability to memorize information from its past inputs, thus influencing its outputs progressively [37]. This mechanism enables RNN to capture context information in the sequential inputs. Due to the *vanishing gradients problem* [38], Long Short-Term

Memory (LSTM) [39] network has been proposed. Both RNN and LSTM are widely used in modeling context for many tasks, such as image classification [40], object detection [41], [42], segmentation [43], [44], scene labeling [45], [46], human re-identification [47] and so on. Besides capturing contextual dependency, LSTM can also selectively allow or block the inputs through the network with its multiplicative gates. In the work of Palangi *et al.* [48], it has been found that LSTM cells can automatically attenuate the unimportant words and detect the salient keywords in the sentence. In the work of Varior *et al.* [47], it also has been shown that LSTM cells can automatically select relevant region parts when comparing two images. Our work aims at building an architecture with an LSTM network as its one component to model bundled object context features for the referring expression tasks. To the best of our knowledge, this is the first attempt to use LSTM networks to model context for referring expressions.

As LSTMs have shown to be powerful models for capturing dependencies among sequential data, many approaches utilize them for non-sequential data. Vinyals *et al.* [49] utilize LSTMs to model dependencies between the elements of a set. Their work demonstrates that for some problems, such as geometric problems, choosing an intuitive ordering of the outputs results in slightly better performance. Jiang *et al.* [50] also use LSTMs to model dependencies between the predicted labels for multi-label image classification. In their work, the label orders are determined according to their occurrence frequencies in the training data. Gkioxari *et al.* [51] predict structured output variables sequentially where the output variables are sorted in descending order according to the detection rates of an unchained feed forward net. Most similar to our work, Chen *et al.* [52] and Liu *et al.* [53] sort the detected objects according to their detected confidence scores to model useful context among them. In this paper, we also investigate several orders that are used to arrange the context objects.

III. Our Model

In this section, we describe our model which utilizes rich information in the image to learn bundled object context for referring expression generation and comprehension, as shown in Fig. 2. It is composed of two modules: a *visual context* LSTM and a *sentence* LSTM. The visual context LSTM receives all the context objects in the image to aggregate a compact and informative context features, mimicking the reaction mechanism of humans - when referring an object in an image, they often perform an exhaustive scan of all the objects. Given the features of the referred object and the learned context features, the sentence LSTM sequentially generates the words of a sentence to describe it.

A. Formulation

We address two associated tasks: referring expression generation and comprehension. The generation task takes an image I and an interior object r as the inputs, and generates an expression s.

$$G: \ I \times r \rightarrow s \tag{1}$$

In order to solve this problem, we design a model $P_G(s|I,\ r)$. With P_G, we have:

$$G(I,\ r) = \underset{s}{argmax}\ P_G(s|I,\ r) \tag{2}$$

To train P_G, we need a dataset with image, object and expression triples $\{(I_i,\ r_i,\ s_i)\}$. Since we will use CNN+LSTM based model to represent $P_G(s|I,\ r)$, of which the core idea is usually to maximize the probability of the description given the input image and object, we can generate the terms of s one after another until the end of it. This is similar to image captioning, except the input is the $\{image,\ object\}$ pair instead of the entire image. The whole model is trained to minimize cross entropy loss which is equivalent to maximize the likelihood:

$$L_{loss} = -\sum\nolimits_{i}^{N}\ \sum\nolimits_{t-1}^{T}(logP_G(s_{i,\ t}|s_{i,\ 1:t-1},\ I_i,\ r_i)) \tag{3}$$

where N is the total number of samples in the training set of the dataset, $s_{i,\ t}$ is the t-th word of the ground truth expression s_i, and T is the length of s_i.

In the comprehension task, given a query q expression and an image I, we are asked to locate an object $\hat{r}$ from a set of objects $R = \{r_i\}$ in the image.

$$C: \ I \times q \rightarrow r,\ r \in R \tag{4}$$

For this task, we can learn a comprehension model which measures the probability of an object given the image and the query expression $P_C(r|I,\ q)$. Given the comprehension model, the selected object is: $\hat{r} = argmax_{r \in R} P_C(r|I,\ q)$. By Bayes' rule, we have

$$P(r|I,\ q) = \frac{P(q|I,\ r)P(r|I)}{\sum_{r' \in R} P(q|I,\ r')P(r'|I)} \tag{5}$$

With the assumption of the uniform prior for $P(r|I)$ [22], the function can be simplified as $\hat{r} = argmax_{r \in R} P_C(q|I,\ r)$. This means that the model which is trained for the generation task can be used to accomplish the comprehension task. So we use the same model for both of these two tasks.

B. The LSTM Model

LSTM networks have been widely used to model contextual dependency and extract discriminative features for image classification, object detection, scene labeling and so on. It has been demonstrated that the LSTM architectures can spot salient keywords from sentences and speech inputs [54], [48]. The work of Varior *et al.* [47] also shows that LSTM networks can automatically select salient strips from images when comparing two images. The internal gating mechanism in the LSTM cells can regulate the propagation of certain relevant context, which enhance the discriminative capability of the features. The core idea of our work is that we use a visual context LSTM network to learn discriminative context from a set of context objects as well as a sentence LSTM network to generate the expression given the target object features and the corresponding context features, as shown in

Fig. 2. Both of the LSTM networks have the same architecture which is described as below.

LSTM takes in the output of the previous time step, as well as the input at the current time step, as the inputs of the current unit. The update equations at time l can be formulated as:

$$i_l = \sigma(\mathbf{W}_{ix}x_l + \mathbf{W}_{im}m_{l-1}) \tag{6}$$
$$f_l = \sigma(\mathbf{W}_{fx}x_l + \mathbf{W}_{fm}m_{l-1}) \tag{7}$$
$$o_l = \sigma(\mathbf{W}_{ox}x_l + \mathbf{W}_{om}m_{l-1}) \tag{8}$$
$$c_l = f_l \odot c_{l-1} + i_l \odot \phi(\mathbf{W}_{cx}x_l + \mathbf{W}_{cm}m_{l-1}) \tag{9}$$
$$m_l = o_l \odot \phi(c_l) \tag{10}$$

where l ranges from the start of the input sequence to the end of it; i_l, f_l and o_l represent the input gate, forget gate, and output gate at time step l respectively; c_l is the state of the memory cell and m_l is the hidden state; $\odot$ represents the element-wise multiplication, $\sigma(\cdot)$ represents the sigmoid function and $\phi(\cdot)$ represents the hyperbolic tangent function; $W_{[\cdot][\cdot]}$ denote the parameters of the model.

C. Modeling Bundled Object Context

Context is an essential part of the input information for referring expression generation and comprehension. A straightforward way to increase the discriminative and informative power of the context features is to leverage rich information in the image. This motivates us to combine all the features of the context objects to form bundled object context with an LSTM network. As the LSTM network can automatically select the relevant context objects and discard the unrelated objects, the combination of all the hidden states can form abundant context features for referring expressions.

We use the visual context LSTM to model bundled object context. It receives the context objects one-by-one and sequentially acquire and aggregate the relevant information, encoding multiple objects features into a compact context vector. To be specific, given an image I, we first detect a fixed number of objects in the image. As the spatial information of objects in an image is vital for modeling object-object relationships and is also an important heuristic for many visual tasks [55], [52], we then sort these objects by their locations in the image. More accurately, they are arranged with the order of from left to right and top to down. Specially, the entire image is also considered as a special object and is arranged at the first position before all of the objects , denoted as R_1. We use $seq(I)$ to denote its initial sequential representations of the image, which contains a sequence of representations $seq(I) = \{R_1, R_2, ...\ R_N\}$, where R_2 to R_N are the local object features extracted from the 'fc7' layer of VGGNet [56] which is pre-trained on ImageNet dataset [57] and R_1 are the global CNN features of the whole image which are also extracted from the 'fc7' layer of ImageNet VGGNet [56]. As a referring expression differentiates one object from other objects, the context sequence can be obtained by replacing the referred one R_i with the whole image R_1, as shown in Fig. 2. It is represented as $seq_{context}(I,\ R_i) = \{R_1, R_2, ...R_1, ...R_N\}$, where R_i is the referred object and is replaced by R_1. In this manner, we can get object-specific conetxt sequence for each target object. Then the visual context LSTM takes in $seq_{context}(I,\ R_i)$ by encoding each object into a fixed length vector. Thus, we have encoding hidden states computed from:

$$h_{t_{en}} = LSTM_{en}(seq_{context}(I,\ R_i)_{t_{en}}, h_{t_{en}-1}),\ t_{en} = 1, 2, ...N \tag{11}$$

Once the hidden representations from all the context objects are obtained, they are combined to obtain the bundled object context $conV_i$ for the referred object R_i as shown below:

$$conV_i = \mathbf{W_C^T}[(h_1)^T, (h_2)^T, ...(h_r)^T, ..., (h_N)^T],\ r = 1, 2, ..., N \tag{12}$$

where $\mathbf{W_C}$ is the transformation matrix we need to learn and $[\cdot]^T$ indicates the transpose operation.

D. Generating Expression

After getting the context features, the sentence LSTM is used to generate the expression describing the target object. It is fed with the features of the referred object along with the bundled context $conV_i$, and then generates the natural language. To represent the object, we use the approach used by Mao *at al.* [22]. Specifically, the features of the referred object are composed of visual appearance representations R_i and location and size representations l_i. R_i are the features extracted from VGGNet, as mentioned in the previous section. l_i are the features encoding the target object location and size. $l_i = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}]$, where x and y are the locations of the top left and bottom right corners of the target object; w and h are the width and height of the object; W and H are the width and height of the image. The final features v_i used to generate the sentence are obtained from the concatenation of these three features.

$$v_i = [R_i,\ conV_i,\ l_i] \tag{13}$$

Given the vector v_i, the language model is trained to minimize the cross entropy, decoding it into its corresponding expression s_i.

$$L(i) = -\sum\nolimits_{t=1}^{T}(log\ p(s_{i,\ t}|s_{i,\ 1:t-1},\ v_i)) \tag{14}$$

E. Training and Inference

The visual context LSTM and the sentence LSTM are optimized end-to-end with the loss in Eq. 14. The overall loss of our model is formally written as:

$$J(\theta) = -\sum\nolimits_{n=1}^{N} log\ p(s_n|r_n, C_n, \theta) \tag{15}$$

where s_n is the expression, r_n is the referred object, C_n is the corresponding context sequence. We also use max-margin loss (also named Maximum Mutual Information training, MMI) [22] to enhance the probability of a referring expression to be high for the true object and low for other objects. For a referring expression, let r_n be the true object and $r_n^{'}$ be a negative object where $r_n^{'} \in R \backslash r_n$, C_n and $C_n^{'}$ be the context sequences corresponding to r_n and $r_n^{'}$ respectively, then the loss function is written as:

$$J'(\theta) = -\sum_{n=1}^{N} log\ p(s_n|r_n, C_n, \theta) + \lambda max(0, M - log\ p(s_n|r_n, C_n, \theta) + log\ p(s_n|r'_n, C'_n, \theta)) \quad (16)$$

To obtain the objects in the image, we train a Faster-RCNN [58] model using the VGG-16 convolutional architecture [56]. The model is first pre-trained on the ImageNet [57] dataset and then is fine-tuned on the validation set of MS COCO [59], as the validation set of MS COCO is also used by Yu *et al.* [25] to train their detectors. Then the model is used to detect objects with high confidence in the image. We train our model which is composed of two LSTM sub-networks on the framework of Caffe [60]. We use the mini-batch stochastic gradient descent method with the batch size of 30. The hidden state size of the visual context LSTM is set to 512, and the hidden size of the language LSTM is set to 1024.

In the testing phase, we run our object detector on the testing image, and the feature representations of the objects are fed into our model. For the comprehension task, given an expression and a set of objects, the object with the highest probability to generate the sentence is selected as the target object. For the generation task, the sentence LSTM generates words sequentially and stops when it generates the special <EOS> word.

IV. Experiments

A. Datasets

We verify the effectiveness of our method on three public datasets: RefCOCO [25], RefCOCO+ [25] and RefCOCOg [22], which provide images with both local objects and corresponding natural language descriptions.

RefCOCO [25] is collected by playing the ReferIt Game [18] on images with two or more objects of the same object category. It is composed of 142,210 referring expressions for 50,000 objects in 19,994 images. We use the splits provided by Yu *et al.* [25]. The training set has 16,994 images, 42,404 objects and 120,624 referring expressions. The validation set has 1,500 images, 3,811 objects and 10,834 referring expressions. The testing partition contains two splits. TestA split has 750 person-centric images, 1,975 objects and 5,657 referring expressions. TestB split contains 750 images, 1,810 objects and 5,095 object-centric referring expressions.

RefCOCO+ [25] is collected in the same way with RefCOCO except its expressions focus more on purely appearance. It has 141,564 expressions for 49,856 objects in 19,992 images. The splits we use is the same with Yu *et al.* [25]. The training set has 16,992 images, 42,278 objects and 120,191 referring expressions. The validation set has 1,500 images, 3,805 objects and 10,758 expressions. TestA split has 750 images, 1,975 objects and 5,726 expressions, and TestB split contains 750 images, 1,798 objects and 4,889 expressions.

RefCOCOg [22] is constructed on Amazon Mechanical Trurk with images containing from 2 to 4 instances of the same object category. It has 25,799 images (without the test set which is not released yet), 49,820 objects and 95,010 referring expressions. We work with the partitioning provided by Mao *et al.* [22]. More precisely, the training partition has 24,968 images, 44,820 objects and 85,474 expressions. The validation partition has 4,650 images, 5,000 objects and 9,536 expressions.

B. Evaluation Metrics and Settings

For the referring expression comprehension task, instead of recursively using the predicted word as the input of the next time step to generate a sentence for each object, the learned model takes the given sentence word by word as inputs. It calculates the per-word cross entropy losses between the given word and the predicted word for each object. Corresponding sentence object pairs would have low average losses, while non-corresponding ones would have higher average losses. The object with the lowest loss is selected as the target one. The evaluation is simply performed by measuring the Intersection over Union (IoU) ratio between a ground truth box and the predicted box for a referring expression, as done in [22], [25], [23]. We use Precision@1, more precisely, if the IoU is larger than 0.5, the prediction is considered as a true positive. Otherwise, we count it as a false positive. We use the average score over all images.

For the referring expression generation task, we evaluate the generated descriptions in the same way as evaluating image caption [7]. We use the most commonly used metric such as BLEU, METEOR and ROUGE. While BELU [61] evaluates a candidate sentence by measuring the fraction of n-grams that appear in a set of references, METEOR [62] evaluates a generated sentence by computing a score based on word level matches between the generation and a set of references. ROUGE [63] counts the number of overlapping units between the generated description and its references.

As referring expressions in RefCOCOg are longer than referring expressions in RefCOCO and RefCOCO+, the max length of the sentences in RefCOCOg is set to 20, and the max length of the sentences in RefCOCO and RefCOCO+ are set to 10. For the max-margin loss, we sample 4 negative examples for each referring expression and its referred object in training. At the same time, the margin M is set to 0.1 and the margin weight λ is set to 1.

C. Describing an Image with Local Objects

We first validate whether *the visual context LSTM* can capture discriminative features for language models to generate descriptions. We propose a framework which use local objects in an image to describe the content of it, as shown in Fig. 3. Different from most existing work where the input image is represented by the CNN features of the whole image, we propose to represent the input image as a sequence of detected objects. These objects are arranged in an order of their locations in the image and fed as the input sequence of the visual context LSTM network. Precisely, the image is represented as $seq(I) = \{R_1, R_2, ...\ R_N\}$, as introduced in Section III-C. In this experiment, the regions used to represent the image are the ones generated by our region proposal network (See Section III-E for details). After the hidden representations from all the objects are obtained, they

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2811621, IEEE Transactions on Multimedia

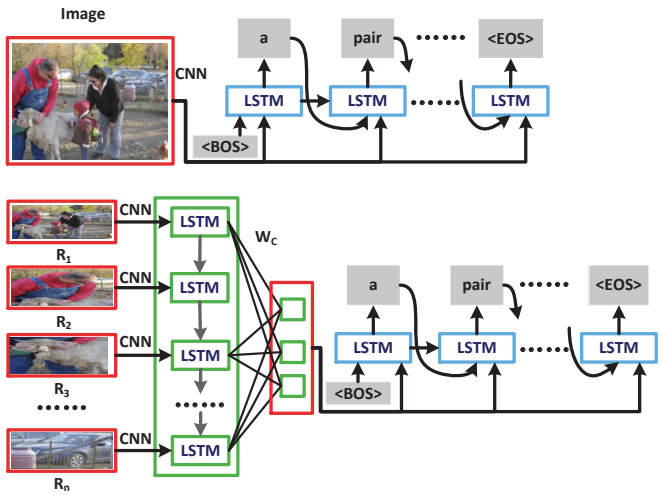


Fig. 3. The overview of describing an image with the entire image and a sequence of objects in the image.

are concatenated with a fully connected layer to form the features representing the image. Then the features are put into a second LSTM to generate natural language to describe the image. To compare the performance of describing an image with the whole image and local objects, we randomly select 15,000 images from the training set of RefCOCOg as our training set and respectively select 1,500 images from the rest as the validation set and testing set.

Fig. 4 demonstrates the performance with various numbers of objects as well as the whole image on our testing set. It can be seen that representing the image with objects is much better than the whole image. The visual context LSTM progressively receives the objects and encodes these local features into a discriminative vector representing the image. With the increase of the objects, more details of the image are revealed, so the performance increases. After reaching a certain threshold, the increase of the objects brings more noise, so the performance decreases. As shown in Fig. 4, when the number of objects is 5 (excluding the first one which is the whole image), the model achieves the best performance.

D. Arranging Context Objects with Different Orders

In order to capture important dependencies among the context objects and at the same time allow for tractable inference, the order of the context objects is an important issue to learn the bundled object context. Besides arranging the context object according to their spatial locations in the image (BOC+MMI-Spatial), we also investigate the other two kinds of orders: the area order (BOC+MMI-Area) and the confidence order (BOC+MMI-Confidence). The area order is that the detected objects are sorted by their areas from big to small, as we usually tend to pay attention to big objects and neglect small objects. The confidence order is that objects are sorted by their confidence scores from big to small, where the confidence scores are provided by the object detection model.

Table I shows the referring expression comprehension results on the RefCOCO dataset when the context objects are arranged with different orders. Both of the models are trained with MMI. With any kind of order, the proposed method with bundled object context (BOC+MMI) works much better than the baseline (MMI) which uses the entire image as the context [22]. More importantly, it also can be observed that the the spatial order achieves the best performance and the other two get comparable performance. So in the following sub-sections, we use the spatial order to arrange the context objects.

TABLE I
REFERRING EXPRESSION COMPREHENSION RESULTS (%) ON REFCOCO WHEN THE CONTEXT OBJECTS ARE ARRANGED WITH DIFFERENT ORDERS

	Test A		Test B	
	GT	DET	GT	DET
BOC+MMI-Spatial	75.89	68.98	76.15	57.21
BOC+MMI-Area	75.62	67.86	75.96	57.02
BOC+MMI-Confidence	-	67.56	-	56.99
MMI [22]	71.72	64.90	71.09	54.51

E. Referring Expression Comprehension

For the referring expression comprehension task, we evaluate the performance of our method on RefCOCO, RefCOCO+ and RefCOCOg. We conduct experiments on two kind of settings as the same in [22], [25], [23], [35]. The first one is that the object proposals are got from ground truth bounding boxes. While the ground truth number of objects in image varies from each other, the proposed method needs a fixed number of objects as its input proposal set, so we select the top k objects with big sizes. The second one is that the object proposals are generated by object detectors. We use the trained Faster-RCNN model to extract k objects with high confidences. When the number of objects is 5, most details in the image are presented and the image level captioning model achieves the best performance (as validated in Section IV-C), so we use $k = 5$ for all three datasets.

The results are shown in Table II. For fair comparisons, we also provide the results that use the same proposals with previous work. Because the split for RefCOCOg in [23] is different from the split (which is used in our paper) in [25], the corresponding results are not presented. It can be observed that our method which uses the learned bundled object context (BOC) works much better than the baseline [22]. For example, when in the RefCOCO dataset, our method not only gets a gain of 4.17 when the object proposals are provided by ground truth bounding boxes (GT), but also has a gain of 4.08 when the proposals are generated from object detectors (DET). At the same time, for the most cases, our method also surpasses the approaches proposed in [23], [25], [35]. We can conclude that the approaches using the same proposals with previous work also get better performance than the baselines. The results demonstrate the effectiveness of our method both in the settings of GT and DET. Fig. 5 shows a few comprehension examples from the test set of RefCOCO. Compared with the baseline [22] which uses the whole image as the context, our model receives information from all the objects in the image and thus pinpoints the referred objects more accurately.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2811621, IEEE Transactions on Multimedia

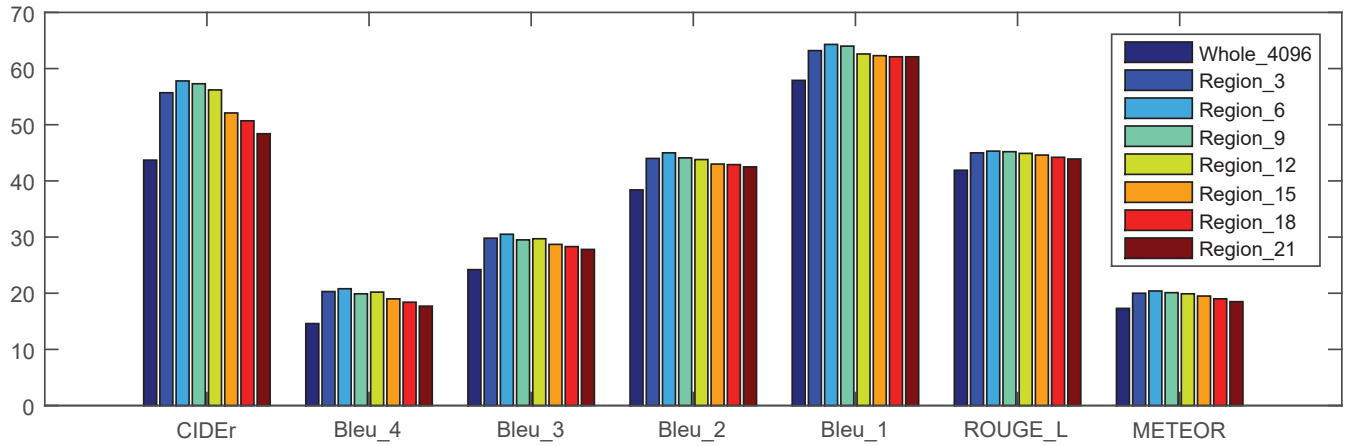


Fig. 4. The effect of number of objects for describing an image. Whole_4096 denotes representing the image with the CNN features of the whole image. Object_3 denotes representing the image with 3 objects within it.

TABLE II
REFERRING EXPRESSION COMPREHENSION RESULTS (%) ON REFCOCO, REFCOCO+ AND REFCOCOG DATASETS. "GT" REPRESENTS THAT THE OBJECT PROPOSALS ARE PROVIDED BY GROUND TRUTH BOUNDING BOXES AND "DET" REPRESENTS THAT THE PROPOSALS ARE GENERATED FROM OBJECT DETECTORS

	RefCOCO				RefCOCO+				RefCOCOg	
	Test A		Test B		Test A		Test B		Val	
	GT	DET	GT	DET	GT	DET	GT	DET	GT	DET
MLE [22]	63.15	58.32	64.21	48.48	48.73	46.86	42.13	34.04	55.16	40.75
MMI [22]	71.72	64.90	71.09	54.51	58.42	54.03	51.23	42.81	62.14	45.85
visdif [25]	67.57	62.50	71.19	50.80	52.44	50.10	47.51	37.48	59.25	41.85
visdif+MMI [25]	73.98	67.64	76.59	55.16	59.17	55.81	55.62	43.43	64.02	46.86
Neg Bag Margin [23]	75.6	58.6	78.0	56.4	-	-	-	-	-	-
Pos & Neg Bag Margin [23]	75.0	58.7	76.1	56.3	-	-	-	-	-	
Com-guided [35]	74.04	68.11	73.43	55.18	60.26	57.05	55.03	43.74	-	-
BOC (Faster-RCNN)	72.18	62.78	73.46	54.19	56.60	48.26	51.90	37.27	61.59	47.79
BOC+MMI (Faster-RCNN)	75.89	68.98	76.15	57.21	61.45	57.27	55.93	43.83	65.75	53.70
BOC+MMI (Multibox)	-	-	-	-	-	-	-	-	-	52.62
BOC+MMI (MCG)	-	66.21	-	57.69	-	-	-	-	-	-
BOC+MMI (Fast-RCNN)	-	68.72	-	57.08	-	57.23	-	43.82	-	-

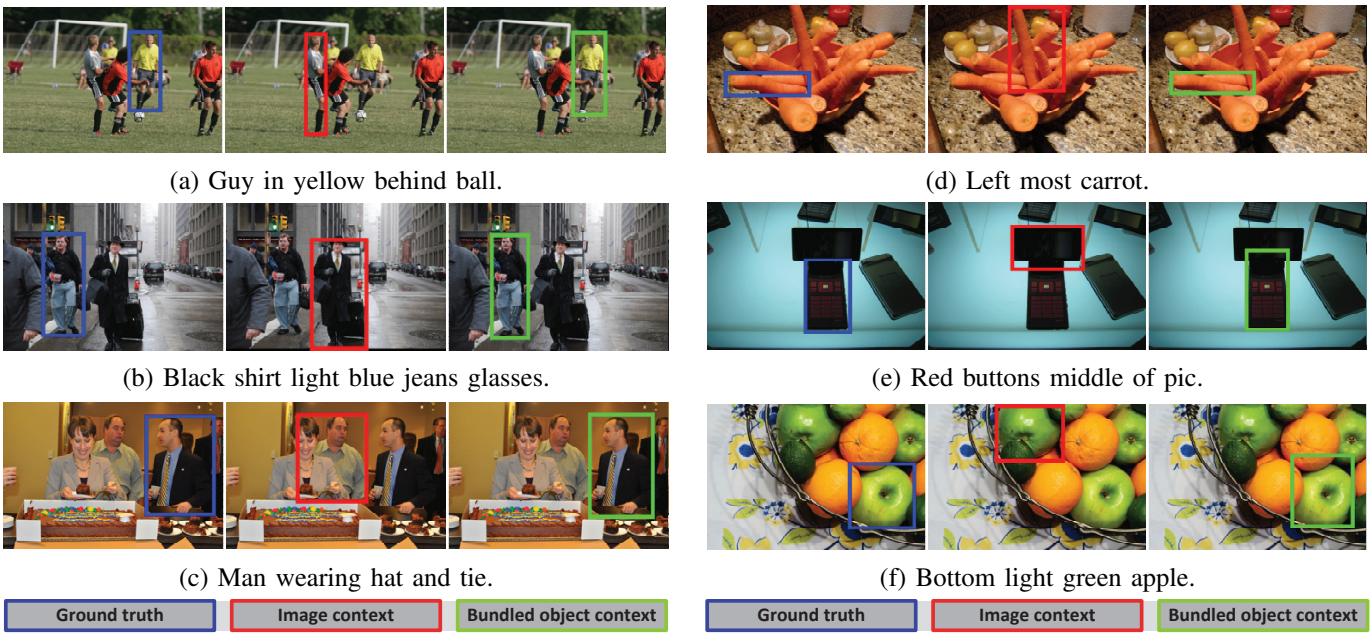


Fig. 5. Referring expression comprehension results on the test set of RefCOCO. (a-c) are examples from the TestA partition, and (d-f) are examples from the TestB partition.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2811621, IEEE Transactions on Multimedia

TABLE III
REFERRING EXPRESSION GENERATION RESULTS ON REFCOCO, REFCOCO+ AND REFCOCOG DATASETS

	RefCOCO Test A				RefCOCO Test B			
	BLEU-1	BLEU-2	ROUGE	METEOR	BLEU-1	BLEU-2	ROUGE	METEOR
MLE [22]	0.477	0.290	0.413	0.173	0.553	0.343	0.499	0.228
MMI [22]	0.478	0.295	0.418	0.175	0.547	0.341	0.497	0.228
visdif [25]	0.505	0.322	0.441	0.184	0.583	0.382	0.530	0.245
visdif+MMI [25]	0.494	0.307	0.441	0.185	0.578	0.375	0.531	0.247
BOC	0.510	0.321	0.427	0.185	0.575	0.376	0.512	0.242
BOC+MMI	0.495	0.317	0.425	0.184	0.579	0.377	0.519	0.249

	RefCOCO+ Test A				RefCOCO+ Test B				RefCOCOg Val			
	BLEU-1	BLU-2	ROUGE	METEOR	BLEU-1	BLEU-2	ROUGE	METEOR	BLEU-1	BLEU-2	ROUGE	METEOR
MLE [22]	0.391	0.218	0.356	0.140	0.331	0.174	0.322	0.135	0.437	0.273	0.363	0.149
MMI [22]	0.370	0.203	0.346	0.136	0.324	0.167	0.320	0.133	0.428	0.263	0.354	0.144
visdif [25]	0.407	0.235	0.363	0.145	0.339	0.177	0.325	0.145	0.442	0.277	0.370	0.151
visdif+MMI [25]	0.386	0.221	0.360	0.142	0.327	0.172	0.325	0.135	0.430	0.262	0.356	0.145
BOC	0.392	0.224	0.348	0.145	0.334	0.174	0.332	0.140	0.443	0.273	0.364	0.151
BOC+MMI	0.401	0.238	0.358	0.153	0.340	0.178	0.339	0.140	0.436	0.270	0.359	0.148

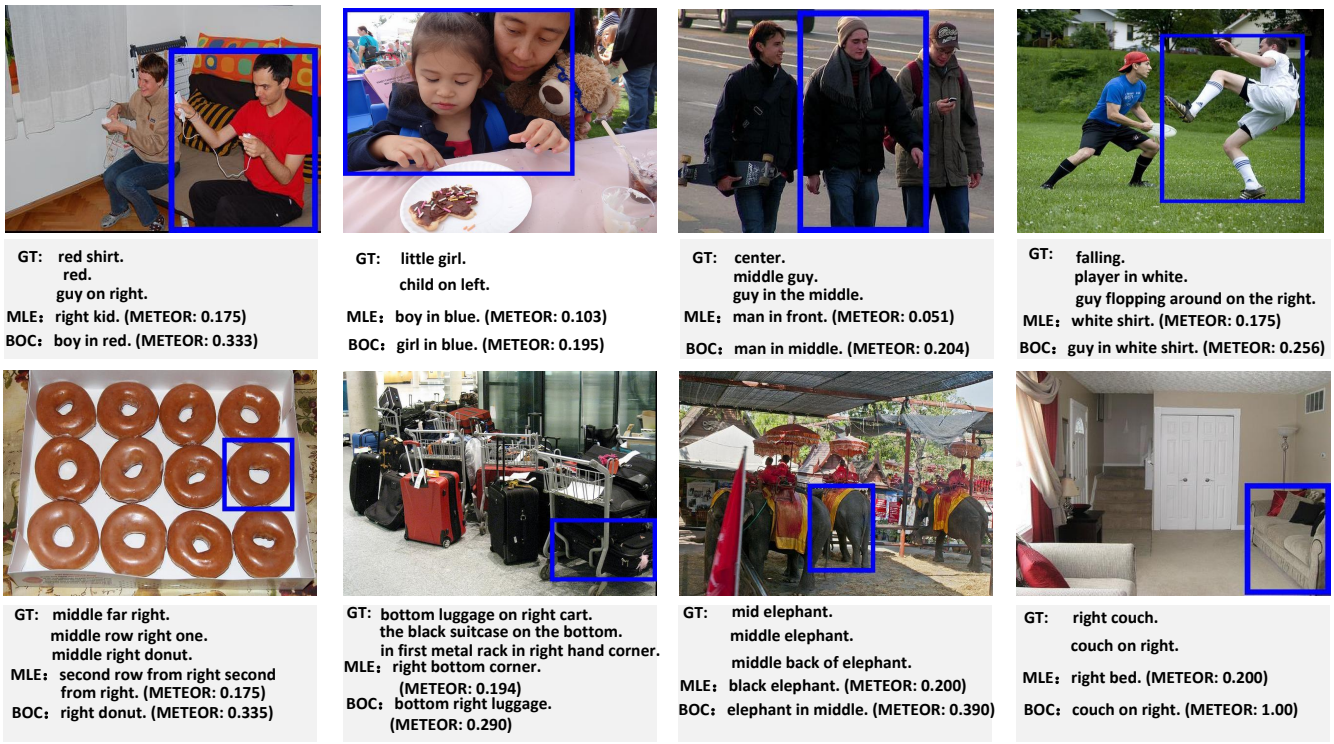


Fig. 6. Referring expression generation results on the test set of RefCOCO. The first row shows two examples in the testA partition and the second shows two example from the testB partition.

F. Referring Expression Generation

For the referring expression generation task, we evaluate the usefulness of the proposed method on RefCOCO, RefCOCO+ and RefCOCO. The baselines are max likelihood (MLE) and Maximum Mutual Information training (MMI) of max likelihood in [22]. We also compare our method with [25] which uses the visual comparison as the context.

The referring expression generation results are shown in Table III. We can observe that our model achieves the best performance on the metric of METEOR. The reason why our method always has better results on this metric is that we use it as the criterion to terminate the training procedure. Meanwhile, our method also gets comparable results on all other metrics. The results show that our method improves generation quality. Unlike the comprehension task, MMI training does not always bring improvement in all the datasets. Fig. 6 shows several referring expression generation results on the test set of RefCOCO. Considering all the context objects in the image, our model captures and aggregates informative context features and thus generates better expressions for the referred objects.

G. Visualization of the Visual Context LSTM

The aim of this experiment is to visualize the properties of the visual context LSTM model and explain how it works in capturing informative and discriminative context information from all of the context objects. We examine the temporal evolution of internal gate states and understand how the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2811621, IEEE Transactions on Multimedia

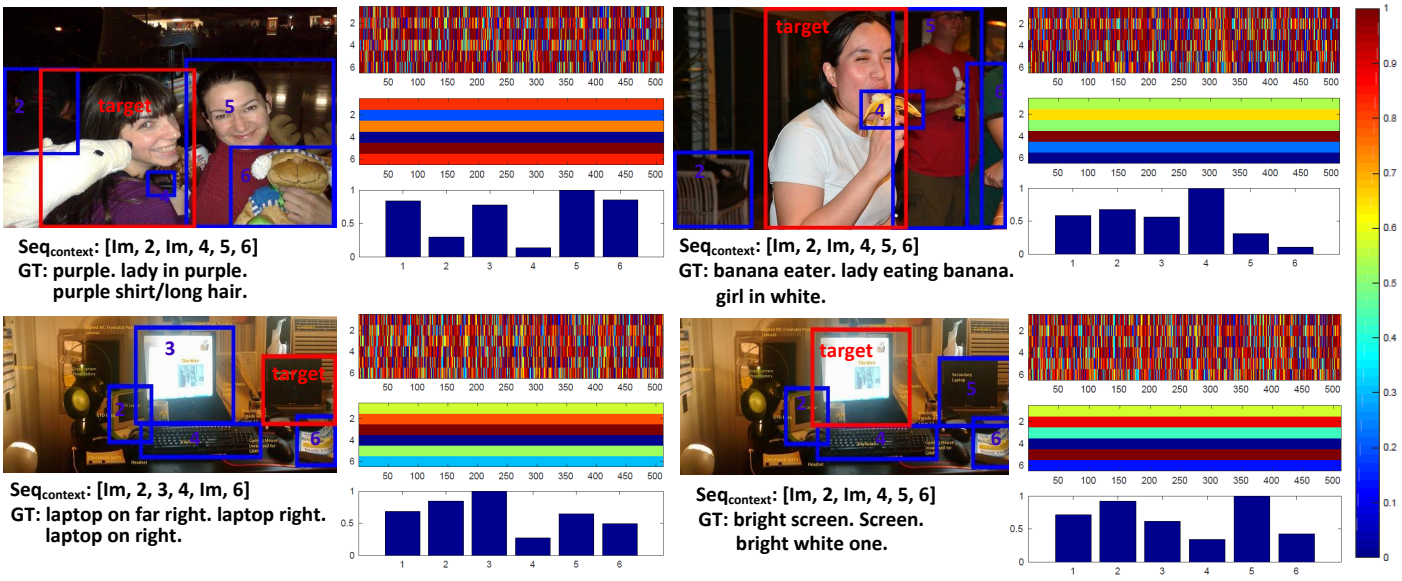


Fig. 7. Visualization of the input gate activation values of the visual context LSTM. The first column shows the images, corresponding referred objects and expressions. The second column shows the original input gate activation values, the mean activation values and the min-max normalization of the mean activation values for each part of the input sequence of the visual context LSTM (Best viewed in color).

visual context LSTM retains valuable context information and attenuates unimportant information.

Fig. 7 shows four examples from the test set of RefCOCO. As the dimension of hidden state of the visual context LSTM is 512, so the gate activation values are 512 dimensional vectors whose values range from 0 to 1. The input, the output and the forget gates of the LSTM architecture modulate whether its memory cell is written to, reset or read from respectively, so the input gate activation values work in such a way that relevant information in the input sequence is propagated and unimportant information is attenuated. From the response (i.e. the input gate activation values) to different objects, we would like to qualitatively answer the question whether the visual context LSTM can select and propagate relevant contextual information or discard the irrelevant contextual information. In Fig. 7, the first column shows the images, corresponding referred objects and expressions. The second column shows the original input gate activation values, the mean activation values and the min-max normalization of the mean activation values for each part of the input sequence of the visual context LSTM. From the left one in the first row, we can observe that to generate the expression "lady eating banana", the visual context LSTM selects and propagates the relevant banana (marked with 4) object as the corresponding normalized mean activation value is big (i.e. 1.0). Meanwhile it attenuates the irrelevant object (marked with 6) which is a part of the body far away from the target object, as the corresponding mean activation value is small (i.e. 0.11). The left one in the second row also demonstrates that other laptops are important context for referring the right laptop while the keyboard is less important. The visualization indicates that meaningful patterns can be learned by the visual context LSTM. By automatically selecting and attenuating context objects, the visual context LSTM can capture and aggregate informative and discriminative context features.

V. Conclusion

In this paper, we have proposed a method for referring expression generation and comprehension, which is composed of a visual LSTM module for learning bundled object context and a sentence LSTM module for generating natural language. The visual context LSTM receives all context objects progressively and encodes all the information to informative and discriminative bundled object context features. The learned bundled object context reveals more details of the image and provides abundant visual information to describe the target object. By analyzing the internal input gate of the visual context LSTM module in our network, we show that it can automatically select the relevant objects and discard the irrelevant ones. Our method is evaluated on several challenging referring expression datasets and it achieves promising results compared to the state-of-the-art. In future work, we will focus on adding reasoning modules to the language model of our framework to generate less ambiguous expressions which contrastively describe referred objects.

References

[1] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.

[2] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363 – 1377, Jul. 2016.

[3] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1201–1216, Jun. 2016.

[4] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[5] X. Jiang, F. Wu, X. Li, Z. Zhao, W. Lu, S. Tang, and Y. Zhuang, "Deep compositional cross-modal learning to rank via local-global alignment," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 69–78.

[6] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891 – 2903, Dec. 2013.

[7] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 3128–3137.

[8] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional lstms," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 988–997.

[9] X. Li, X. Song, L. Herranz, Y. Zhu, and S. Jiang, "Image captioning with both object and scene information," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1107–1110.

[10] J. Krishnamurthy and T. Kollar, "Jointly learning to parse and perceive: Connecting natural language to the physical world," *Trans. Associat. Comput. Linguist.*, vol. 1, pp. 193–206, 2013.

[11] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," in *Proc. 19th Int. Conf. Mach. Learn.*, 2012, pp. 1671–1678.

[12] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: image caption with region-based attention and scene factorization," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/1506.06272

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural image caption generation with visual attention," in *Proc. 32th Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[14] R. Lebret, P. O. Pinheiro, and R. Collobert, "Pharse-based image captioning," in *Proc. 32th Int. Conf. Mach. Learn.*, 2015, pp. 2085–2094.

[15] X. Jia, S. Gavves, B. Fernando, and T. Tuytelaars, "Guiding long-short term memory for image caption generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2407–2415.

[16] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Dec. 2016, pp. 4565–4574.

[17] A. Rohrbach, M. Rohrbach, and R. Hu, "Grounding of textual phrases in images by reconstruction," in *Proc. Eur. Conf. on Comput. Vis.*, 2016, pp. 817–834.

[18] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proc. EMNLP*, 2014, pp. 787–798.

[19] E. Krahmer and K. van Deemter, "Computational generation of referring expressions: A survey," *Comput. Linguist.*, vol. 38, no. 1, pp. 173–218, 2012.

[20] N. FitzGerald, Y. Artzi, and L. Zettlemoyer, "Learning distributions over logical forms for referring expression generation," in *Proc. EMNLP*, 2013, pp. 1914–1925.

[21] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A joint speaker-listener-reinforcer model for referring expressions," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1612.09542

[22] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 11–20.

[23] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expressing understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 792–807.

[24] M. Mitchell, K. van Deemter, and E. Reiter, "Generating expressions that refer to visible objects," in *Proc. 13th Conf. North American Chapter Associat. Comput. Linguist.: Human Languag. Technol.*, June 2013, pp. 1174–1184.

[25] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–85.

[26] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.

[27] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (M-RNN)," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1412.6632

[28] 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 Recog.*, Jun. 2016, pp. 203–212.

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

[30] C. Liu, J. Mao, F. Sha, and A. L. Yuille, "Attention correctness in neural image captioning," in *Proc. Assoc. Advan. Artific. Intel.*, 2017, pp. 4176–4182.

[31] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 4555–4564.

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

[33] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik., "Phrase localization and visual relationship detection with comprehensive linguistic cues," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1611.06641

[34] T. Winograd, "Understanding natural language," *Cognit. Psychol.*, vol. 3, no. 1, pp. 1–191, 1972.

[35] R. Luo and G. Shakhnarovich, "Comprehension-guided referring expressions," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1701.03439

[36] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1611.09978

[37] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 2012, vol. 385.

[38] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Network.*, vol. 5, no. 2, pp. 157–166, Mar 1994.

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

[40] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, and B. Wang, "Learning contextual dependence with convolutional hierarchical recurrent neural networks," *IEEE Trans. Imag. Proc.*, vol. 25, no. 7, pp. 2983–2996, Jul. 2016.

[41] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-Outside net: Detecting objects in context with skip pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 2874–2883.

[42] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May. 2017.

[43] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 108–124.

[44] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph LSTM," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 125–143.

[45] B. Shuai, Z. Zuo, G. Wang, and B. Wang, "DAG-Recurrent neural networks for scene labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 3620–3629.

[46] W. Byeon, T. M. Breuel, F. Raue, and M. R. Liwicki, "Scene labeling with LSTM recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 3547–3555.

[47] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamses long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 135–153.

[48] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Trans. Audio, Speech, Languag. Proces.*, vol. 24, no. 4, pp. 694 – 707, Apr. 2016.

[49] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," in *Proc. Int. Conf. Learn. Repres.*, 2016.

[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 Recog.*, Jun. 2016, pp. 2285–2294.

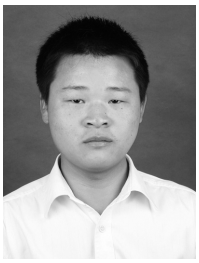
[51] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *Proc. Europ. Conf. Comput. Vision*, 2016, pp. 728–743.

[52] X. Chen and A. Gupta, "Spatial memory for context reasoning in object detection," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1704.04224

[53] C. Liu, F. Sun, C. Wang, F. Wang, and A. Yuille, "MAT: A multimodal attentive translator for image captioning," in *Proc. Assoc. Advan. Artific. Intel.*, 2017.

[54] S. Fernández, A. Graves, and J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *Proc. 17th Int. Conf. Artif. Neural Network.*, Berlin, Heidelberg, 2007, pp. 220–229.

[55] H. Xu and K. Saenko, "Ask, attend and answer: Exlporing question-guided spatial attention for visual question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 451–466.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556
[57] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. Jour. Comput. Vis.*, vol. 115, no. 3, pp. 211 – 252, 2015.
[58] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.
[59] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common object in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[60] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[61] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, 2002, pp. 311–318.
[62] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2th Workshop Statist. Mach. Transl. ACL*, 2007, pp. 228–231.
[63] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. of the ACL-04 Workshop*, vol. 8, 2004, pp. 74–81.



Xiangyang Li received the B.E. degree from the Wuhan Institute of Technology, Wuhan, China, in 2012, the M.E. degree from the Capital Normal University, Beijing, China, in 2015. He is currently working toward the Ph.D. degree at the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

His research interests include large-scale image classification, joint learning with language and vision, computer vision, and pattern recognition.

Shuqiang Jiang (SM'08) is a professor with the Institute of Computing Technology, Chinese Academy of Sciences(CAS) and a professor in University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. His research interests include multimedia processing and intelligent understanding, pattern recognition, and computer vision. He has authored or coauthored more than 100 papers on the related research topics. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, NSFC Excellent Young Scientists Fund in 2013, Young top-notch talent of Ten Thousand Talent Program in 2014. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012.

Prof. Jiang is the senior member of IEEE and CCF, member of ACM, Associate Editor of IEEE Multimedia, Multimedia Tools and Applications. He is the vice Chair of IEEE CASS Beijing Chapter, vice chair of ACM SIGMM China chapter. He has served as an organization member of more than 20 academic conferneces, including the general chair of ICIMCS 2015, program chair of ICIMCS2010, grand chanllenge chair of ACM Mutlimedia 2018, special session chair of ACM ICMR2018 and PCM2008, etc. He has also served as a TPC member for many conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, ICME, ICIP, etc.