Natural Language Object Retrieval

Ronghang Hu[1] Huazhe Xu[2] Marcus Rohrbach[1,3] Jiashi Feng[4] Kate Saenko[5] Trevor Darrell[1]
[1]University of California, Berkeley [2]Tsinghua University [3]ICSI, Berkeley
[4]National University of Singapore [5]University of Massachusetts, Lowell
{ronghang, rohrbach, trevor}@eecs.berkeley.edu, xhz12@mails.tsinghua.edu.cn
elefjia@nus.edu.sg, saenko@cs.uml.edu

Abstract

In this paper, we address the task of natural language object retrieval, to localize a target object within a given image based on a natural language query of the object. Natural language object retrieval differs from text-based image retrieval task as it involves spatial information about objects within the scene and global scene context. To address this issue, we propose a novel Spatial Context Recurrent ConvNet (SCRC) model as scoring function on candidate boxes for object retrieval, integrating spatial configurations and global scene-level contextual information into the network. Our model processes query text, local image descriptors, spatial configurations and global context features through a recurrent network, outputs the probability of the query text conditioned on each candidate box as a score for the box, and can transfer visual-linguistic knowledge from image captioning domain to our task. Experimental results demonstrate that our method effectively utilizes both local and global information, outperforming previous baseline methods significantly on different datasets and scenarios, and can exploit large scale vision and language datasets for knowledge transfer.

1. Introduction

Significant progress has been made in object detection in recent years; with the help of Convolutional Neural Networks (CNNs), it is possible to detect a predefined set of object categories with high accuracy [8, 7], and the number of categories in object detection has grown over 10K to 100K with the help of domain adaptation [12] and hashing [2]. However, in practical application scenarios, instead of using a predefined fixed set of object categories, one would often prefer to refer to an object with natural language rather than use a predefined category label. Such natural language query can include different types of phrases such as categories, attributes, spatial configurations and interactions with other objects, such as "the young lady in a white dress sitting on the left" or "white car on the right" in Figure 1.

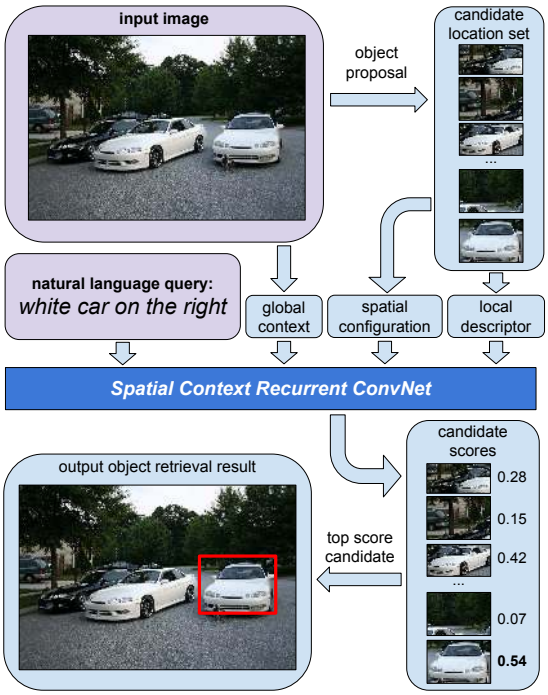

Figure 1. Overview of our method. Given an input image, a text query and a set of candidate locations (e.g. from object proposal methods), a recurrent neural network model is used to score candidate locations based on local descriptors, spatial configurations and global context. The highest scoring candidate is retrieved.

In this paper, we address the problem of natural language object retrieval: given an image and a natural language description of an object as query, we want to retrieve the object by localizing the object in the image. Natural language object retrieval can be seen as a generalization of generic object detection and has a wide range of applications, such as handling natural language commands in robotics where the user may ask to a robot to pick up "the TV remote control on the shelf".

We frame natural language object retrieval as a retrieval task on a set of candidate locations in a given image in this paper, as shown in Figure 1, where candidate locations can come from object proposal methods such as EdgeBox [33]. We observe that simply applying text-based image retrieval systems on the image regions cropped from candidate locations for this task leads to inferior performance, as natural language object retrieval involves spatial configurations of objects and the global scene as context. For example, to decide how likely an object in a scene corresponds to "the man in a blue jacket sitting on the right in front of the house", one needs to look at both the object to determine whether it is "the man" (category), "in blue jacket" (attribute) and "sitting" (action), and its spatial configuration within the scene to determine whether it is "on the right", and the whole image as global contextual information to determine whether it is "in front of the house". Although both text-based image retrieval and natural language object retrieval involve jointly modeling images and text, they are different vision and language domains with domain shift from whole images to bounding boxes.

To address these issues, we propose the Spatial Context Recurrent ConvNet (SCRC) model to learn a scoring function that takes the text query, candidate regions, their spatial configurations and global context as input and outputs scores for candidate regions. Inspired by the Long-term Recurrent Convolutional Network (LRCN) [4], an effective recurrent architecture for both image captioning and image retrieval, we use a two-layer LSTM network structure where the embedded text sequence and visual features serve as input to the first layer and the second layer, respectively. However, we note that it is possible to build our model on other recurrent network architectures such as [25, 31].

Compared with other types of visual-linguistic models such as bag-of-words [27], one of the advantages of using a recurrent neural network as scoring function is that the whole model can be easily learned end-to-end via simple back propagation, allowing visual feature extraction and text sequence embedding to be adapted to each other, and we show that it significantly outperforms a previous method using bag-of-words. Another advantage is that it is easy to utilize relatively large scale image-text datasets from other domains like image captioning (e.g. MSCOCO [23]) to learn a vision-language model, by first pretraining the model on the image captioning task, and then adapting it to natural language object retrieval task through fine-tuning. One of the main challenges for natural language object retrieval is the lack of large scale datasets with annotated object bounding box and description pairs. To address this issue, we show that it allows us to transfer visual-linguistic knowledge learned from the former task to the latter one by first pretraining on the image caption domain and then adapting it to the natural language object retrieval domain. This pretraining and adaptation procedure improves the performance and avoids over-fitting, especially when the object retrieval training dataset is small.

2. Related work

Natural language object retrieval. Based on a bag of words sentence model and embeddings derived from ImageNET classifiers, [10] addresses a similar problem as ours and localizes an object within an image based on a text query. Given a set of candidate object regions, [10] generates text from those candidates represented as bag-of-words using category names predicted from a large scale pretrained classifier and compares the word bags to the query text. Other methods generate visual features from query text and match them to image regions, e.g. through a text-based image search engine [1] or learn a joint embedding of text phrases and visual features. Concurrent with our work, [24] also proposes a recurrent network model to localize objects from given descriptions.

Grounding Objects from Image Descriptions. Given an image and its description sentence, [18] aligns sentence fragments to image regions by embedding the detection results from a pretrained object detector and the dependency tree from a parser with a ranking loss. [17] builds on [18] and replaces the dependency tree with a bidirectional RNN. Canonical Correlation Analysis (CCA) is used in [26] to learn a joint embedding of image regions and text snippets to localize each object mentioned in the caption. [22] uses a structure prediction model to align text to image and reasons about object co-reference in text for 3D scene parsing. Concurrent with this paper, [28] uses an attention model to ground referential phrases in image descriptions by attending to regions where the phrases can be best reconstructed.

Image Captioning. Image captioning methods take an input image and generate a text caption describing it. Recently, methods based on recurrent neural networks [32, 31, 25, 4] have shown to be effective on this task. LRCN [4] is one of these recent successful methods and involves a two-layer LSTM network with the embedded word sequence and image features as input at each time step. We use LRCN as our base network architecture in this work and incorporate spatial configurations and global context into the recurrent model for natural language object retrieval.

Image Retrieval. Text-based image retrieval systems select from a set of images an image that best matches the query text. In image retrieval, a ranking function is learned through a recurrent neural network [25, 4], metric learning [13], correlation analysis [21] and other methods [6, 20]. It was shown in [4] that a probabilistic image captioning model such as LRCN can also be used as an image retriever by using the probability of the query text sequence conditioned on the image $p(S_{query}|I)$ generated by image captioning model as a score for retrieval.

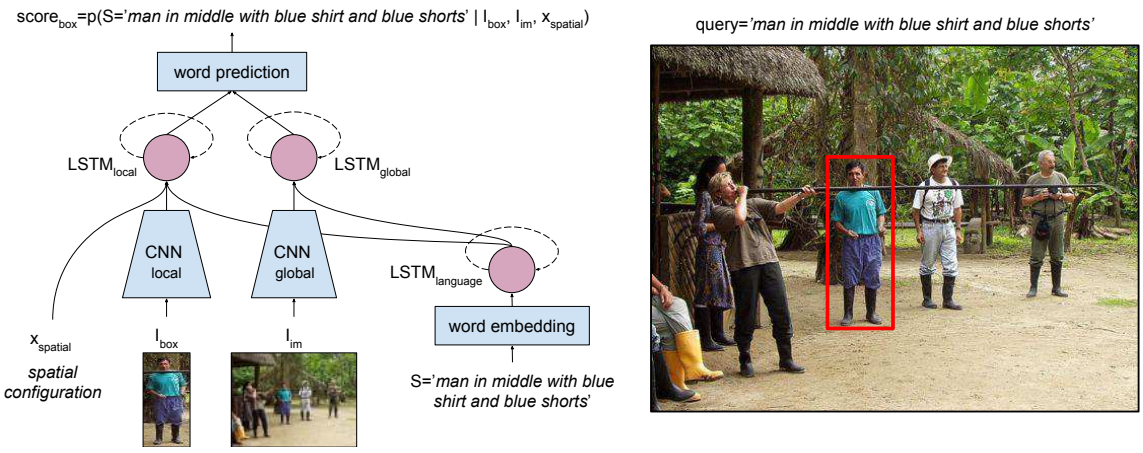

Figure 2. Our Spatial Context Recurrent ConvNet (SCRC) for natural language object retrieval. The recurrent network in our model contains three LSTM units. Two CNN's are used to extract local image descriptors and global scene-level contextual feature respectively. Parameters in word embedding, word prediction and three LSTM units are initialized by pretraining on image captioning dataset.

3. Our model

In this section, we describe our Spatial Context Recurrent ConvNet (SCRC) model for natural language object retrieval and the training procedure in details. At test time, an image, a natural language object query and a set of candidate bounding boxes (e.g. from object proposal methods such as EdgeBox [33]) are provided. The system needs to select from the candidate set a subset of bounding boxes that match the query text.

3.1. Spatial Context Recurrent ConvNet

Inspired by the architecture of LRCN [4], our Spatial Context Recurrent ConvNet (SCRC) model for natural language object retrieval consists of several components as illustrated in Figure 2. The model has three Long Short-Term Memory (LSTM) [11] units denoted by $\text{LSTM}_{language}$, LSTM_{local} and LSTM_{global}, a local and a global Convolutional Neural Network (CNN), a word embedding layer and a word prediction layer. At test time, given an image I, a query text sequence S and a set of candidate bounding boxes $\{b_i\}$ in I, the network outputs a score s_i for the i-th candidate box b_i based on local image descriptors x_{box} on b_i, spatial configuration $x_{spatial}$ of the box with respect to the scene, and global contextual feature $x_{context}$.

In this work, the local descriptor x_{box} is extracted by CNN_{local} from local region I_{box} on b_i, and we use feature extracted by another network CNN_{global} on the whole image I_{im} as scene-level contextual feature $x_{context}$. The spatial configuration of b_i is an 8-dimensional representation

$$x_{spatial} = [x_{\min}, y_{\min}, x_{\max}, y_{\max}, x_{\text{center}}, y_{\text{center}}, w_{box}, h_{box}] \quad (1)$$

where w_{box} and h_{box} are the width and height of b_i. We normalize image height and width to be 2 and place the origin at the image center, so that coordinates range from -1 to 1.

The words $\{w_t\}$ in the query text sequence S are represented as one-hot vectors and embedded through a linear word embedding matrix as Ew_t, and processed by $\text{LSTM}_{language}$ as the input time sequence. At each time step t, LSTM_{local} takes in $[h^{(t)}_{language}, x_{box}, x_{spatial}]$ (concatenation of the three vectors, where $h^{(t)}_{language}$ is the hidden state from $\text{LSTM}_{language}$), and LSTM_{global} takes in $[h^{(t)}_{language}, x_{context}]$. Finally, based on $h^{(t)}_{local}$ and $h^{(t)}_{global}$, a word prediction layer predicts the conditional probability distribution of the next word based on local image region I_{box}, whole image I_{im}, spatial configuration $x_{spatial}$ and all previous words it have seen so far, as

$$\begin{aligned} & p(w_{t+1}|w_t, \cdots, w_1, I_{box}, I_{im}, x_{spatial}) \\ = \; & \text{Softmax}(W_{local} h^{(t)}_{local} + W_{global} h^{(t)}_{global} + r) \end{aligned} \quad (2)$$

where W_{local} and W_{global} are weight matrices for word prediction and r is a bias vector. $\text{Softmax}(\cdot)$ is a softmax function over a vector to output a probability distribution.

We note that when setting $W_{local} = 0$ in Eqn. 2, our Spatial Context Recurrent ConvNet (SCRC) model is equivalent to the LRCN model [4] for image captioning and image retrieval by only modeling $p(S|I_{im})$ to predict a text sequence S based on the whole image I_{im} while ignoring I_{box} and $x_{spatial}$. This makes it possible to pretrain the model on the image captioning in Section 3.2 to obtain a good parameter initialization for visual-linguistic modeling, and transfer knowledge from large image captioning datasets.

We use VGG-16 net [30] trained on ILSVRC-2012 dataset [29] as the CNN architecture for CNN_{local} and CNN_{global} and extract 1000-dimensional *fc8* outputs as

x_{box} and $x_{context}$, and use the same LSTM implementation as in [4], where the gates are computed as

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) && (3)\\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) && (4)\\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) && (5)\\
g_t &= \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) && (6)
\end{aligned}
$$

All the three LSTM units have 1000-dimensional state h_t.

At test time, given an input image I, a query text S and a set of candidate bounding boxes $\{b_i\}$, the query text S is scored on i-th candidate box using the likelihood of S conditioned on the local image region, the whole image and the spatial configuration of the box, computed as

$$
\begin{aligned}
s &= p(S|I_{box}, I_{im}, x_{spatial})\\
&= \prod_{w_t \in S} p(w_t|w_{t-1}, \cdots, w_1, I_{box}, I_{im}, x_{spatial}) \quad (7)
\end{aligned}
$$

and the highest scoring candidate boxes are retrieved.

3.2. Knowledge transfer from image captioning

To exploit paired image-text data in image captioning datasets, and to learn a good initialization of parameters in word embedding, word prediction and three LSTM units, we first pretrain our model on an image captioning dataset, by restricting $W_{local} = 0$ in Eqn. 2, which is equivalent to training a LRCN model [4]. We follow the procedure in [4] for pretraining on image captioning. During pretraining, the probability of ground truth image caption $p(S_{gt}|I_{im})$ is maximized over the training image-sentence pairs, and the whole network is optimized with standard Stochastic Gradient Descent (SGD). We refer to [4] for the training details on image captioning.

Since we restrict $W_{local} = 0$ in Eqn. 2 during pretraining, the parameters in LSTM_{local} are not learned. To obtain a good initialization of this unit, we copy those weights in Eqn. 3 – 6 from LSTM_{global} to LSTM_{local}. The weights over the extra 8 dimensions of $x_{spatial}$ are initialized with zero. We also copy W_{global} to W_{local} to initialize word prediction weights.

After pretraining on the image captioning task, the parameters in our model already encode useful knowledge of word embedding and decoding and sequence prediction based on image features. The knowledge is transferred to the natural language object retrieval task in Section 3.3.

3.3. Training for object retrieval

After pretraining, we adapt the SCRC model to natural language object retrieval. In this paper, we assume that the training dataset consists of N images, with each image containing M_i ($i = 1, \cdots, N$) annotated objects, and each object annotated by a bounding box and $K_{i,j}$ ($i = 1, \cdots, N$, $j = 1, \cdots, M_i$) text descriptions (an object can be described more than once with different descriptions). At training time, each instance is an image-bounding box-description tuple $(I_i, b_{i,j}, S_{i,j,k})$, where I_i is the whole image, $b_{i,j} = [x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ is the bounding box of the j-th object and $S_{i,j,k}$ is a description text in natural language such as "the black and white cat".

Our model for natural language object retrieval can be trained via maximizing the probability of the object description text in ground truth annotations conditioned on the local image region I_{box} and the whole image I_{im} as context, which is analogous to training a generic object detection system. Many state-of-the-art generic object detectors [8, 7] are built by turning object detection into a classification problem on candidate bounding boxes produced either from a sliding window or an object proposal mechanism, and a classifier is trained by maximizing the probability of ground truth object category label. In natural language object retrieval, the description text of an object can be seen as a generalized "label" of the object, and maximizing its conditional probability is similar to training a "generalized classifier" whose output is a sequence of word labels rather than a single category label.

Given a natural language object retrieval dataset, we construct all tuples $(I_i, b_{i,j}, S_{i,j,k})$ from the ground truth annotations as training instances (multiple tuples are constructed if there are multiple descriptions for the same object). For each annotated object in the training set, an image patch I_{box} is cropped from the whole image I_{im} based on bounding box of that object region, with its spatial configuration $x_{spatial}$ constructed through Eqn. 1. We define the loss function during training as

$$
L = -\sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=1}^{K_{i,j}} \log(p(S_{i,j,k}|I_{box\,i,j}, I_{im\,i}, x_{spatial_{i,j}})) \quad (8)
$$

where N is the number of images, M_i is the number of annotated objects in i-th image, $K_{i,j}$ is the number of natural language descriptions associated with the j-th object in that image, and $p(S_{i,j,k}|I_{box\,i,j}, I_{im\,i}, x_{spatial_{i,j}})$ is computed by Eqn. 7.

During training, the model parameters are initialized from the pretrained network in Section 3.2, and fine-tuned using SGD with a smaller learning rate, allowing the network to adapt to natural language object retrieval domain. The whole network is trained end-to-end via back propagation. Our model is implemented using Caffe [16] and our code and data are available at http://ronghanghu.com/text_obj_retrieval.

4. Experiments

We evaluate our method on different datasets from small scale to relatively large scale. More experimental results

can be found in the supplementary material or in [15]. Since [10] solves a similar problem to our paper, we adopt it as our baseline. In [10], a large scale fine-grained classifier of 7K object classes is trained on ImageNET [3]. Each box in the candidate set is classified into one of the 7K classes, and a bag of words is extracted from the predicted object class based on its ImageNET [3] synset containing category name and synonyms. Then, the word bag is projected to a vector space, and matched to the projected query text using cosine distance to obtain a score. The sentence projection (embedding) in [10] is predefined and the only training involved in training the 7K object classifier. Note that [10] also proposes an instance match model that relies on online APIs at test time. As in this work we assume a self-contained system without resorting to other APIs, we only use the category model (CAFFE-7K) in [10] as our baseline.

As our recurrent architecture is inspired by LRCN [4], which is shown to be effective for both image captioning and image retrieval, we also compare our model to LRCN. We use the LRCN model trained on MSCOCO [23] for image captioning task as an object retriever by evaluating it on candidate bounding boxes. Given an image I with a set of candidate boxes and a query text S_{query}, we compute $p(S_{query}|I_{box})$, the probability of the query text S_{query} conditioned on the local image region I_{box} outputted by LRCN as a score for each box in the candidate set, and retrieve highest scoring candidates.

4.1. Object retrieval evaluation on ReferIt dataset

The ReferIt dataset [19] is the biggest publicly available dataset containing image regions with descriptions at the time of writing. It contains 20,000 images from IAPR TC-12 dataset [9], together with segmented image regions from SAIAPR-12 dataset [5], and 120K annotated descriptions for the image regions collected in a two-player game that aims to make the image region identifiable from the annotation. The ReferIt dataset also contains some ambiguous (e.g. "anywhere") and mistakenly annotated examples where the annotation does not correspond to any object. To evaluate on this dataset, we split the 20,000 images (together with their annotations) into 10,000 for training and validation and 10,000 for test, and construct image-bounding box-description tuples on all annotated image regions as training instances. There are 59,976 *(image, bounding box, description)* tuples in the trainval set and 60,105 in the test set. In our experiments on this dataset, we only use the bounding boxes of annotated regions during training and evaluation. The bounding boxes are obtained from the segmentation regions in SAIAPR-12 dataset corresponding to the clicks by annotators. Note that although [19] introduces the ReferIt dataset, it does not propose a baseline method for object retrieval based on text query.

As described in Section 3, we first pretrain a SCRC model on MSCOCO dataset [23] for image captioning. The training details such as hyper-parameters of SGD follow [4]. After pretraining, we copy the weights in LSTM and the word prediction layer to the local part of the network as mentioned in Section 3.2. Then the pretrained SCRC model is adapted to the natural language object retrieval task following the procedure in Section 3.3. The model is fine-tuned on image-bounding box-description tuples in ReferIt trainval set with back propagation.

Ablations. To test the effect of incorporating spatial configurations $x_{spatial}$ and scene-level contextual feature $x_{context}$, we evaluate different setups during fine-tuning on ReferIt. By setting $x_{spatial}$ and W_{global} to 0 during fine-tuning and testing, the model can only learn to score a box based on local image descriptors x_{box} from candidate boxes, denoted by **SCRC (w/o context, spatial)**. Similarly, by setting W_{global} to 0, the model can learn a scoring function on x_{box} and $x_{spatial}$ but cannot utilize scene-level context, denoted by **SCRC (w/o context)**.

As a comparison, we directly trained a SCRC model on ReferIt without first pretraining on MSCOCO, and set $x_{spatial}$ and W_{global} to 0 during training and testing, denoted by **SCRC (w/o context, spatial, transfer)**. The CNN parameters in the model are initialized from VGG-16 net [30] and other parameters are randomly initialized. In all the training above, the whole SCRC model is trained end-to-end with SGD, allowing visual feature extraction and textual sequence prediction to be optimized jointly.

At test time, all the 4 SCRC models mentioned above, the bag-of-words model (CAFFE-7K) in [10] and LRCN [4] as an object retriever on candidate boxes are compared on the ReferIt test set. The LRCN model is trained on MSCOCO dataset for image captioning as described in [4] to learn a probabilistic generative model $p(S|I)$, and we use it to score a candidate region I_{box} based on a text query S_{query} by computing the probability of the text conditioned on the local region, i.e. $p(S_{query}|I_{box})$ as a baseline.

We evaluate with two testing scenarios: In the first scenario similar to the experiment in [10], given an image and a text query, the model is asked to retrieve the corresponding image region from all annotated regions in that image. In the second scenario, which is a harder task but closer to real applications, given a text query the model retrieves an image region from a set of candidate bounding boxes produced by object proposal methods. A retrieved region is considered as correct if it overlaps with ground truth bounding box by at least 50% IoU. In this experiment, we use top 100 proposals from EdgeBox [33] as our candidate box set.

Results. Table 1 shows the top-1 precision (the percentage of the highest-scoring region being correct) in the first scenario where the candidate set is all annotated boxes in the image. Note that CAFFE-7K cannot return informative results when none of the words in query are in its category

Method	P@1-NR	P@1
CAFFE-7K [10]	32.53%	27.73%
LRCN [4]	-	38.38%
SCRC (w/o context, spatial, transfer)	-	61.03%
SCRC (w/o context, spatial)	-	64.09%
SCRC (w/o context)	-	70.15%
SCRC	-	**72.74%**

Table 1. Top-1 precision of our method compared with baselines on annotated bounding boxes in ReferIt dataset. See Section 4.1 for details.

Method	R@1	R@10
CAFFE-7K [10]	10.38%	26.20%
LRCN [4]	8.59%	31.86%
SCRC (w/o context, spatial, transfer)	14.53%	40.72%
SCRC (w/o context, spatial)	15.78%	42.54%
SCRC (w/o context)	17.68%	44.77%
SCRC	**17.93%**	**45.27%**
Oracle	59.38%	59.38%

Table 2. Performance of our method compared with baselines on 100 EdgeBox proposals in ReferIt dataset. See Section 4.1 for details.

names (leading to an empty bag and same score for all regions), whereas our SCRC model can always return deterministic result since it can represent unknown words with "<unk>". Similar to [10], we evaluate with "P@1-NR" corresponding to non-random top-1 precision computed on the those informative results and "P@1" corresponding to top-1 precision on all cases including non-informative results, where random guess is used. Results show that our full SCRC model achieves the highest top-1 precision. In Table 1, it can be seen that pretraining on image captioning, adding spatial configuration, and adding scene-level context all improve the performance, with adding spatial configuration $x_{spatial}$ leading to the most significant performance boost. This is no surprise, as spatial configuration not only benefits in cases where spatial relationship is directly involved in the query (e.g. "the man on the left"), but also enables the network to learn a prior distribution of object locations (e.g. "sky" is usually at the top of the scene while "ground" is usually at the bottom).

Table 2 shows the result of the second scenario on 100 EdgeBox proposals, where "R@1" is the recall of the highest scoring box (the percentage of the highest scoring box being correct), and "R@10" is the percentage of at least one of the 10 highest scoring proposals being correct. We also report "Oracle" (or equivalently "R@100"), the percentage of at least one of all 100 proposals being correct, as an upper-bound of all object retrieval systems in this scenario. It can be seen that results in Table 2 follow the same trend as in Table 1, with our full SCRC model achieving the highest recall. Figure 4 shows examples of correctly retrieved objects at top-1 using 100 EdgeBox proposals, where the highest scoring candidate region from our SCRC model overlaps with ground truth annotation by at least 50% IoU, and Figure 5 shows some failure cases, where retrieved top-1 candidate region fails to match ground truth.

By comparing "SCRC (w/o context, spatial)" and "SCRC (w/o context, spatial, transfer)" in Table 1 and Table 2, it can also be seen that the pretraining and adaptation procedure described in Section 3 outperforms directly training on retrieval dataset, showing that pretraining allows the model to transfer useful visual-linguistic knowledge from image captioning dataset.

Also, our SCRC model outperforms the bag-of-words CAFFE-7K model and LRCN model significantly. Compared with our model, CAFFE-7K method suffers from information loss by first projecting image region to category names and limited vocabulary drawn from predefined object category names, and is not end-to-end trainable. Although LRCN model trained on MSCOCO for image captioning task is effective for text-based image retrieval as shown in [4], directly running it as an object retriever on a set of candidate boxes results in inferior performance. This is because object retrieval and image retrieval are different domains, and LRCN model as a object retriever does not encode spatial configuration or global context.

4.2. Object retrieval evaluation on Kitchen dataset

We also evaluate and compare our method with the baseline model [10] on the same Kitchen dataset as used in [10]. Kitchen is a dataset with 606 images sampled from the kitchen/household sub-tree of ImageNET hierarchy [3], with 10 different descriptions annotated for each image. Since objects in this dataset almost occupy the entire images, instead of using retrievals on candidate object proposals boxes, in [10] the performance of the object retrieval is evaluated at image-level. During testing, for each query text, the candidate set consists of 11 images with ground truth and 10 distractors. The distractors are sampled either from the same Kitchen dataset ("Kitchen" experiment) or from the whole ImageNET ("ImageNET" experiment), with the latter being an easier task. Performance of object retrieval is evaluated using top-1 precision.

To evaluate our method on this dataset, we split the dataset into two parts, with 300 images as trainval set and 306 images as test set. Similar to Section 4.1, we first pretrain a SCRC model on MSCOCO dataset [23] for image captioning, and then fine-tune the model on the trainval set. The our model is tested through image-level retrieval on the candidate set of ground truth and 10 distractors, where we use the feature extracted from the entire image as x_{box}. Since the dataset involves no spatial configurations or scene-level contextual information, we set $x_{spatial}$ and W_{global} in Eqn. 2 to zero during fine-tuning and testing, so

Figure 3. Correctly retrieved examples in Kitchen dataset, where the highest scoring object (green) matches ground truth.

the model can only learn to score a candidate based x_{box}. As this dataset is a much smaller than ReferIt, we observe that transferring knowledge from MSCOCO significantly boosts the performance and avoids overfitting.

Results. Table 3 shows the top-1 precision (P@1) of our method together with the baseline on the test set. The first column "Kitchen" corresponds to sampling the 10 distractors from the same Kitchen dataset, while the second column corresponds to sampling distractors from the whole ImageNET 7K dataset [3]. Similar to Section 4.1, **LRCN** refers to directly running LRCN model on the candidate images as a retriever. **SCRC (w/o context, spatial, transfer)** refers to the SCRC model directly trained on the trainval part of the Kitchen dataset, with convolutional layer initialized from VGG-16 net, and LSTM unit, word embedding and word prediction weights randomly initialized. **SCRC (w/o context, spatial)** corresponds to first pretraining on MSCOCO and then fine-tuning on Kitchen trainval set as described in Section 3. As the dataset contains no spatial configuration or scene-level context information, we cannot test our full SCRC model on it. It can be seen from Table 3 that in both scenarios, pretraining on image captioning and fine-tune on natural language object retrieval leads to the best performance, outperforming the baseline bag-of-words model CAFFE-7K and LRCN. Figure 3 shows some correctly retrieved object examples from Kitchen dataset, where the highest scoring candidate matches the ground truth. Both the ground truth and the 10 distractor images are sampled from the same Kitchen dataset in Figure 3.

Moreover, as Kitchen dataset has only 606 objects and is more than 100 times smaller than ReferIt dataset, "SCRC (w/o context, spatial)" has significantly higher accuracy than "SCRC (w/o context, spatial, transfer)". This shows that pretraining on MSCOCO for image captioning dataset improves the performance of natural language object retrieval significantly on this relatively smaller dataset, by transferring the visual-linguistic knowledge from the former task to the latter task. As a reference, we note that [10] also uses an instance model and achieves higher overall performance. The instance model sends the query and candidate image regions to online APIs such as Google Image Search and FreeBase on the fly at test time. As in this work we assume a self-contained system that can be applied without resorting to Internet APIs on the fly, we only compare with the category model CAFFE-7K in [10].

Method	Kitchen	ImageNet
CAFFE-7K [10]	51.34%	57.50%
LRCN [4]	40.35%	63.22%
SCRC (w/o context, spatial, transfer)	54.02%	74.08%
SCRC (w/o context, spatial)	**61.62%**	**81.15%**

Table 3. Performance of different methods on the Kitchen dataset. See Section 4.2 for details.

5. Conclusion

In this paper, we address natural language object retrieval with Spatial Context Recurrent ConvNet (SCRC), a recurrent neural network model that scores a candidate box based on local image descriptors, spatial configurations and global scene-level context. We show that incorporation of spatial configuration and global context improves the performance of natural language object retrieval significantly. The recurrent network model used in our method leads to an end-to-end trainable scoring function, which significantly outperforms baseline methods.

Also, we demonstrate that natural language object retrieval can benefit from transferring knowledge learned on image captioning through pretraining and adaptation. As one of the difficulties for natural language object retrieval systems is the lack of large datasets with object-level annotation, we show that this problem can be alleviated by exploiting datasets with image-level annotations, which are often easier to collect than object-level descriptions. As follow up to this work we show successful results by encoding the phrase rather than scoring it [28] and also predicting image segmentations instead of bounding boxes [14].

Acknowledgments. M. Rohrbach was supported by a fellowship within the FITweltweit-Program of the German Academic Exchange Service (DAAD). J. Feng was supported by NUS startup grant R263000C08133. This work was supported by DARPA, AFRL, DoD MURI award N000141110688, NSF awards IIS-1427425 and IIS-1212798, and the Berkeley Vision and Learning Center.

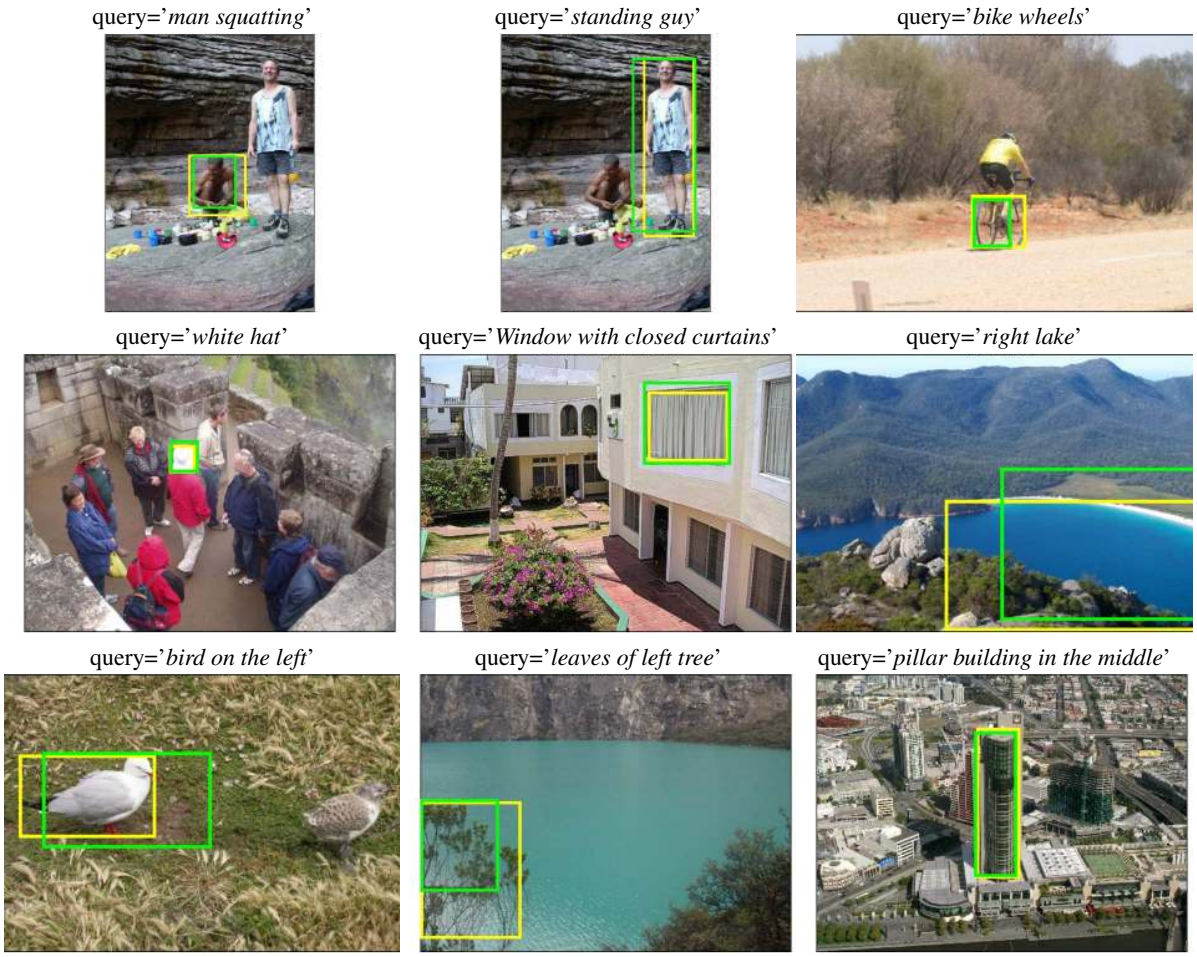

Figure 4. Correctly localized examples (IoU ≥ 0.5) on ReferIt with EdgeBox. Ground truth in yellow and correctly retrieved box in green.

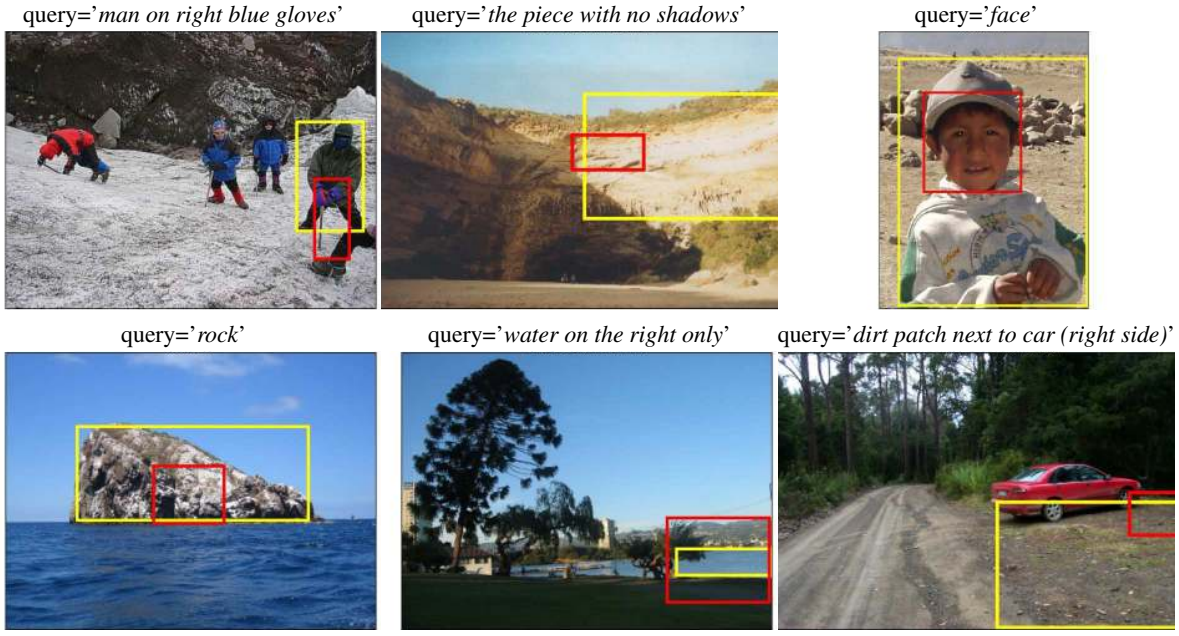

Figure 5. Failure cases (IoU < 0.5) on ReferIt with EdgeBox. Ground truth in yellow and incorrectly retrieved box in red. Some failures cases are caused by ambiguity of the query and some due to wrong annotations in the dataset.

References

[1] R. Arandjelovic and A. Zisserman. Multiple queries for large scale specific object retrieval. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–11, 2012. 2

[2] T. Dean, M. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, J. Yagnik, et al. Fast, accurate detection of 100,000 object classes on a single machine. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1814–1821. IEEE, 2013. 1

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 5, 6, 7

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015. 2, 3, 4, 5, 6, 7

[5] H. J. Escalante, C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428, 2010. 5

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013. 2

[7] R. Girshick. Fast R-CNN. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 4

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014. 1, 4

[9] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, pages 13–23, 2006. 5

[10] S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell. Open-vocabulary object retrieval. In *Robotics: Science and Systems*, 2014. 2, 5, 6, 7

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[12] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544, 2014. 1

[13] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2072–2078. IEEE, 2006. 2

[14] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. *arXiv preprint arXiv:1603.06180*, 2016. 7

[15] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. *arXiv preprint arXiv:1511.04164*, 2015. 5

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, volume 2, page 4, 2014. 4

[17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[18] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2

[19] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014. 5

[20] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics (TACL)*, 2015. 2

[21] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014. 2

[22] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3558–3565. IEEE, 2014. 2

[23] 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 *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014. 2, 5, 6

[24] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016. 2

[25] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proceedings of the International Conference on Learning Representations*, 2015. 2

[26] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[27] M. Ren, R. Kiros, and R. Zemel. Image question answering: A visual semantic embedding model and a new dataset. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[28] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. *arXiv preprint arXiv:1511.03745*, 2015. 2, 7

[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. *International Journal of Computer Vision*, pages 1–42, 2014. 3

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 5

[31] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 2

[32] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the International Conference on Machine Learning (ICML)*, 2015. 2

[33] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 391–405. Springer, 2014. 2, 3, 5