Parallel Attention: A Unified Framework for Visual Object Discovery through Dialogs and Queries

Bohan Zhuang[*], Qi Wu[*], Chunhua Shen[†], Ian Reid, Anton van den Hengel
University of Adelaide, Australia; Australian Centre for Robotic Vision
{bohan.zhuang,qi.wu01,chunhua.shen,ian.reid,anton.vandenhengel}@adelaide.edu.au

Abstract

Recognising objects according to a pre-defined fixed set of class labels has been well studied in the Computer Vision. There are a great many practical applications where the subjects that may be of interest are not known beforehand, or so easily delineated, however. In many of these cases natural language dialog is a natural way to specify the subject of interest, and the task achieving this capability (a.k.a, Referring Expression Comprehension) has recently attracted attention. To this end we propose a unified framework, the ParalleL AttentioN (PLAN) network, to discover the object in an image that is being referred to in variable length natural expression descriptions, from short phrases query to long multi-round dialogs. The PLAN network has two attention mechanisms that relate parts of the expressions to both the global visual content and also directly to object candidates. Furthermore, the attention mechanisms are recurrent, making the referring process visualizable and explainable. The attended information from these dual sources are combined to reason about the referred object. These two attention mechanisms can be trained in parallel and we find the combined system outperforms the state-of-art on several benchmarked datasets with different length language input, such as RefCOCO, RefCOCO+ and GuessWhat?!.

1. Introduction

Despite the fact that Object Detection has become a figurehead challenge in Computer Vision, the number of applications of the technology is limited by the fact that it demands a pre-defined set of class labels and large number of pre-prepared training images for each. This means not only that the list of objects of interest is fixed, and determined long in advance, but also that their appearance must remain fixed. The limitations of this approach are visible in the rise of mediation strategies such as Domain Adaptation [9, 22], Transfer Learning [28, 32], Zero-Shot
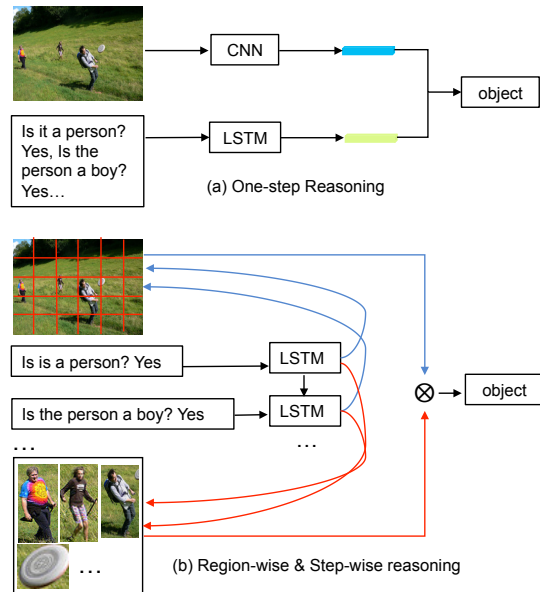

Figure 1: One-step reasoning vs. our proposed region-wise&step-wise reasoning framework for the task of referring expression. In conventional frameworks (a), visual and language features are embedded into a joint space for one-step reasoning. However, in our approach (b), we propose to recurrently discover the target with a parallel attention mechanism, with region and step-wised query.

learning [31, 46], and a subset of the Meta-Learning approaches [34]. More fundamentally, the traditional Machine Learning approach whereby the problem and training data are specified before the solution is devised, and then trained, inherently implies that every instance of the target class is of interest, in every image it processes.

A more flexible, and practically applicable approach is to define the subject of interest at test time, not least because it allows instance-level detection. It also supports the specification of subjects of interest that were not possible at training time. Better even than a system that allows test-time specification of the subject of interest, however, is one that will interactively cooperate towards an accurate detec-

---

[*]First two authors contributed equally to this work.

[†]C. Shen is the corresponding author.

tion, particularly in the presence of ambiguity. This is the challenge we consider here.

Referring expressions are used frequently by people to identify or indicate particular objects within their physical environment [44]. The length of the expressions can range from a very short phase (2-3 words) to a multi-round dialog (such as the GuessWhat?! game [6]). The longer the expression is, the more information is provided, however, the harder the problem is because more details need to be analysed and more steps of reasoning are required. Previous works [6, 43, 44] have primarily appled a CNN-RNN pipeline that uses a CNN to encode the image content and an LSTM to encode the expression. The encoded features are then jointly embedded and used to to locate the object that is related to the expression, as shown in the Figure 1 (a). This mechanism works well when the expression is short and the number of the proposed objects is limited. However, when the expression is too long, in dialog form, or there are too many potential objects in the image, using the global representation to encode the expression and image into a single vector fails because such a one-step process does not have the flexibility required to relate multiple parts of the expression to multiple parts of the image.

Instead, we argue that the representation of expression should be stepwise and the representation of image should be region-wise. More significantly, we argue that the referring process should be also carried out stepwise, *i.e.* after a stepwise representation (such as a word in a sentence or a question-answer pair in a dialog) is given, it should first focus on some potential regions, with the expression or dialog being 'listened to' continually, the number of potential regions decreases and the area of potential interest becomes smaller, until it converges to a single region, as shown in the figure 1 (b). For example, given an expression 'the woman in the middle wearing brown jacket', regions corresponding to all of the women in the image are considered first. After reading 'in the middle', regions of 'woman' at the border of the image are eliminated. Finally, once the input 'wearing brown jacket' is given, only one region is left.

On the basis of the above we propose a unified framework, the ParalleL AttentioN (PLAN) network, which is the main contribution of this paper, to recurrently discover the object in an image that is being referred to natural language descriptions, from short phrases query to long multi-round dialogs. Specifically, a two-way attention mechanism is applied to localize the referred-to region in the global contextual features from the whole image (*i.e.*, image-level attention) and to select the referred-to proposal region (*i.e.*, proposal-level attention) from a set of such region proposals. The 'image-level' attention can help identify referring expressions that are related to the contextual information, such as 'the girl outside the cinema'. The 'cinema' here is then the global scene context we need to consider in the

reasoning process. The 'proposal-level' attention learns to weight different candidate object proposals, when the referring expression is associated with multiple objects, such as 'the girl at the left side of the fire hydrant', then the model will learn to focus on the object proposal of 'girl' and the 'fire hydrant', but not the global scene. The attention mechanism proposed here allows the model to sequentially 'listen to' the referring expression so that a step-wise reasoning process is achieved. We evaluate our PLAN model on what is currently the largest referring expression dataset, the ReferCOCO, ReferCOCO+ [43] and ReferCOCOg [26], which has a different length of the input expression. Our model outperforms the previous state-of-art by a large margin, for example, our single model even outperforms an ensemble model (and with Reinforcement Learning) on a test split on the ReferCOCO. We further evaluate our model on the recently released GuessWhat?! [6] dataset, which requires an agent to point out the object in an image that is being discussed by a 'Questioner' and an 'Oracle' via multiple rounds of dialog. Here our model also outperforms the previous state-of-art significantly.

As a side contribution, because our model uses a recurrent process to discover the object, we can display the referring updating process along with the input expression query and the dialog, which makes the multiple steps of reasoning in the Refer Expression Comprehension visualizable and explainable. Qualitative results (see Figure 3) show that we can produce excellent visualization performance.

2. Related work

Referring Expressions The are two distinct challenges relating to Referring Expressions: generation and comprehension. The generation task requires a model to generate a language expression for the given region, which is very similar to the dense image captioning task [15]. Referring expression comprehension aims to localize the regions being described by a given referring expression [13,14,26,27,43]. Given a set of extracted candidate regions, each region is scored by the model with respect to the referring expression and the region with the highest score is selected as the final grounding result. In this paper, we mainly focus on improving the referring expression comprehension task rather than the generation task.

Several works [14, 26] propose to use local visual features or global image features as the feature representation. To better employ the structural information between different candidates, some works [13, 27, 43] further explicitly incorporating modeling context between objects into referring expression. In [13, 27], the models are proposed to handle inter-object relationships for grounding a referential expression into a pair of regions. Yu *et al*. [43] argues that visual appearance comparison can help localize the referred object by looking into other surrounding candidates. Luo *et*
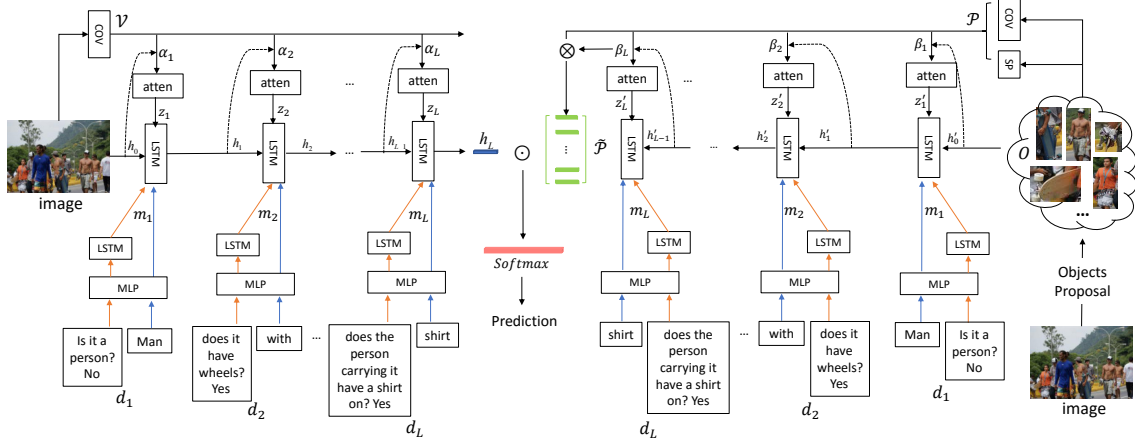
Figure 2: The proposed ParalleL AttentioN (PLAN) framework for recurrent object discovery. The left-side is the image-level framework that is used to encode the entire image information and the referring language descriptions by recurrently attending on different image regions. The right side is the proposal-level framework with the proposal regions and referring expressions as input. It is designed to recurrently attend on proposals based on the language information and mutually encode both visual and textual representations. The both side representations are further combined to effectively localize the referring target. An additional LSTM encoder (orange arrow) is added if the expression unit d_i is a Question-Answer pair, otherwise, we directly input the word into a MLP, followed by an LSTM.

al. [25] propose to utilize models trained for comprehension task to generate better expressions. Different from these methods, we propose to incorporate both contextual image features and local visual features in a unified framework, the both features are weighted by a parallel attention mechanism, which make us recurrently discover the object that referred by the expression.

Phrase grounding Phrase grounding aims to localize the objects described in the phrase. The main problem is to learn the correlation between visual and language descriptions. To solve this problem, Karpathy *et al.* [17] propose to align the objects and the fragments of language into the embedding space with a structured max-margin objective and [16] further replace the dependency tree of the language parser with a bidirectional RNN. Some approaches [30, 37] use a Canonical Correlation Analysis [11] based method to learn the correlation between visual and language modalities. Recently, Hu *et al.* [14] proposed an SCRC model to integrate spatial configuration and global scene-level contextual information into the network.

The Attention Mechanism An attention mechanism was first successfully introduced within the image captioning task by [40]. Based on this, Lu *et al.* [24] further proposed a co-attention model for VQA that jointly reasons about language and the image. Then in [23], an adaptive attention model with a visual sentinel was applied to decide when to attend. You *et al.* [42] run a set of attribute detectors to get a list of visual attributes and fuse them into the RNN hidden state. Rather than statically attending to the image using the given expressions, we instead propose to recurrently update the attention weight with the stepwise expression query or

the dialog. Also we propose a parallel attention mechanism for embedding global and local features in a unified framework.

Vision and Language Our work is part of a group of recent methods combining vision and language. Recent work in this area includes a variety of approaches to image captioning [7, 15, 19, 20, 36, 38, 40, 41], visual question answering (VQA) [2, 8, 12, 24, 39, 45, 47] and visual relationship detection [48, 49]. Image captioning seeks to generate a natural language description for the whole given image, while the VQA requires the agent to answer previously unseen visual questions about an image. Most recently, a new vision-and-language task, Visual Dialog [4–6], demands that an agent participate intelligently in a dialog about an image. In our work, we extend the phase/sentence based referring expression task to dialogs, *i.e.*, tackling the "Guesser" task proposed in [6] for identifying a specific object in an image through a sequence of questions and answers.

3. The PLAN Model

In this section, we describe our unified model that takes as input an image and a set of natural language expressions and outputs a bounding box that contains the object that is referred by the expressions. Formally, we denote an input image as I containing a set of N object proposals $\mathcal{O} = \{O_1, O_2, ..., O_N\}$. For the GuessWhat benchmark only, each object candidate is assigned an object category $c_i \in \{1, ..., C\}$ where C is the number of object categories. The set of referring expressions is $\mathcal{D} = \{d_1, ...d_L\}$, where d_i is either a question-answer pair or a single word, depending on whether the input is dialog or a sentence description,

and L is total number of expressions. The output of the proposed model is a probability distribution over the object proposals and the region with the highest probability is selected as the grounding result.

The core of our proposed model relies on a ParalleL AttentioN (PLAN) model, which takes the encoded image features, language expressions features and the candidate proposal object features as input. On one side, we use word level (for phase and sentence input) or question-answer pair level (for dialog input) features to guide the image attention, *i.e.*, to attend to the regions that are correlated with the expression. At the other side, we use the language features to recurrently weight the object proposasl until all the expressions have been 'listened to'. The parallel attended features are combined to reason about the target object.

In the following sections, we first describe the feature encoding for the different inputs in Sec. 3.1 and explain how to recurrently discover the target object with parallel attention mechanism in Sec. 3.2. The implementation details are described in Sec. 3.3 and the whole framework is illustrated in Figure 2.

3.1. Feature encoding

The global visual feature To encode the full image while maintaining the spatial information therein, we first rescale the image to 224x224 and then pass it through a VGG-16 network [35] pre-trained on ImageNet to obtain its $Conv5\_3$ feature, denoted as $\mathcal{V} = \{\mathbf{v}_1, ..., \mathbf{v}_K\}$, where $K = 49$.

Object proposal features For each candidate object proposal O_i, the corresponding feature is composed of two parts, the first is the CNN feature u_i extracted as described for the global visual features, the $Conv5\_3$ feature from the VGG-16 for each object proposal. To further increase the expressive power, similar to [6, 13, 14, 26, 43, 44], we also embed the spatial representations of the proposal regions. Following [6, 44], the spatial information of the bounding box of each object is encoded as the an 8-d vector: $s_i = [x_{\min}, y_{\min}, x_{\max}, y_{\max}, x_{center}, y_{center}, w_{box}, h_{box}]$, where w_{box} and h_{box} are width and height of the bounding box. The image height and width are normalized to the range $[-1, 1]$ and the center of the image is set as the origin. For the experiments on the GuessWhat?! [6] dataset, the additional object category information for each candidate is also used, for a fair comparison with the previous state-off-the-art. We denote the feature representation as $\mathcal{P} = \{\mathbf{p}_1, ..., \mathbf{p}_N\}$, where N is the number of object proposals and each $\mathbf{p}_i = [u_i; s_i; c_i]$, where c_i is the object category label for the i-th object[1].

Referring expression features Since we argue that the representation of the expression should be processed step-

---
[1]c_i is only used in GuessWhat?! experiment for a fair comparison. We additionally evaluate the model without this object category information.

wise, in contrast to previous work that encodes the whole expression into a single vector, we represent each expression unit d_i separately. Each d_i, is first tokenized by the word tokenizer from the nltk toolkit [3]. Each word is represented as an one-hot vector in a pre-constructed dictionary $\mathbb{D}$. Then we convert each one-hot vector into a dense embedding with a learnable MLP layer. If the expression is a dialog (multiple rounds of question-answer pairs), we pass the expression unit d_i (a QA pair as a whole) to an LSTM encoder and take the last hidden state as the representation (as shown in Figure 2 by the orange arrow). If the expression is a single sentence, the output of the MLP for each word will be directly used as the representation for the expression unit d_i (as shown in the Figure 2 by the blue arrow). For simplicity, we define $\mathcal{M} = \{\mathbf{m}_1, ..., \mathbf{m}_L\}$ as the encoded representation for the expression unit d_i, for either case.

3.2. Recurrent parallel attention mechanism

Given the encoded visual feature representations and a sequence of encoded language descriptions, the next step is to utilize these effectively to identify the correct target object. One naive approach is to concatenate all the features as a vector representation and then dot-product with the features of candidate objects one-by-one, followed by a softmax to obtain a distribution over the objects. The object with the highest probability will be treated as the referring target. However, in this approach, the language and the visual features are encoded separately without learning from each other. Ideally, the regions that are concentrated on should be updated along with the referring expression being 'listened to'. To solve this problem, we propose to recurrently discover the object by sequentially 'listening to' the referring expressions. What's more, a parallel attention mechanism is applied to localize the referring region within the global contextual features from the whole image (*i.e.* image-level) and to select the referent proposal (*i.e.* proposal-level).

3.2.1 Image-level attention

Image-level attention corresponds to the left portion of Figure 2, within which we use expression features to attend over the global image. We use an LSTM to encode the visual feature and the language feature simultaneously. We define $\mathbf{h}_t$ as the hidden state at time t as:

$$\mathbf{h}_t = LSTM(\mathbf{m}_t, \mathbf{z}_t, \mathbf{h}_{t-1}) \tag{1}$$

where $\mathbf{m}_t$ is the feature representation of the t-th expression unit (that is $\mathbf{m}_i$ above, where $i = t$), $\mathbf{z}_t$ is the attended image features at time t. Note that the index of the language description corresponds to the time step in LSTM encoding. We compute $\mathbf{z}_t$ with an attention mechanism, with the convolutional image feature $\mathcal{V} = \{\mathbf{v}_1, ..., \mathbf{v}_K\}$ and $\mathbf{h}_{t-1}$,

which is the hidden state of the LSTM at time $t - 1$. We feed them through a MLP layer separately followed by a softmax function to get the attention distribution over the spatial location on $\mathcal{V}$:

$$\begin{aligned}\mathbf{e}_{ti} &= \tanh\left(\mathbf{W}_v\mathbf{v}_i + \mathbf{W}_h\mathbf{h}_{t-1}\right), \\ \boldsymbol{\alpha}_{ti} &= softmax(\mathbf{e}_{ti}),\end{aligned} \quad (2)$$

where $\mathbf{W}_v$ and $\mathbf{W}_h$ are the embedding matrices and $\boldsymbol{\alpha}_{ti}$ is the attention weights over the convolutional feature $\mathbf{v}_i$ at time step t. Then the attended image feature can be obtained by:

$$\mathbf{z}_t = \sum_{i=1}^{K} \boldsymbol{\alpha}_{ti}\mathbf{v}_i \quad (3)$$

We thus use the previously encoded language and image information $\mathbf{h}_{t-1}$ as context to attend on the image. And, in return, we utilize the attended image feature $\mathbf{z}_t$ as context together with the language representation $\mathbf{m}_t$ to predict the next hidden state of the LSTM. With this strategy, we can recurrently encode different image information given different referring language descriptions, and the last hidden state ($\mathbf{h}_{t=L}$) of the LSTM is taken as the final representation of the image-level attention, which will be used in the final referring step.

3.2.2 Proposal-level attention

The proposal-level attention corresponds to the right-hand side of Figure 2. In contrast to the image-level attention that applies over the densely divided regions, the proposal-level attention model focuses on only the proposal regions that contain the candidate objects. The input is thus a set of proposal regions and a sequence of language descriptions, and the model should learn to recurrently attend to different candidates with varying language description input. Similar to the image-level attention, the hidden state at time t can be represented as:

$$\mathbf{h}_t' = LSTM(\mathbf{m}_t, \mathbf{z}_t', \mathbf{h}_{t-1}') \quad (4)$$

where $\mathbf{z}_t'$ is the summation of the attended proposal features. Given the last hidden state $\mathbf{h}_{t-1}'$ of the LSTM and the proposal container $\mathbf{P} = \{\mathbf{p}_1, ..., \mathbf{p}_N\}$, we derive the attention score for each proposal as follows:

$$\begin{aligned}\mathbf{e}_{ti}' &= \tanh\left(\mathbf{W}_p\mathbf{p}_i + \mathbf{W}_h'\mathbf{h}_{t-1}'\right), \\ \boldsymbol{\beta}_{ti} &= softmax(\mathbf{e}_{ti}'),\end{aligned} \quad (5)$$

where $\boldsymbol{\beta}_{ti}$ is the attention score for proposal $\mathbf{p}_i$ at time t, and $\mathbf{z}_t'$ is computed by:

$$\mathbf{z}_t' = \sum_{i=1}^{N} \boldsymbol{\beta}_{ti}\mathbf{p}_i \quad (6)$$

In contrast to the image-level attention that uses the final hidden state to represent the information, for the proposal-level attention, we use the learned proposal attention weights at the last time step (when the last expression

d_L is 'listened to') $\boldsymbol{\beta}_{Li}$ to weight each of the proposal candidate objects features $\mathbf{p}_i$, to obtain the attended features for each object proposal:

$$\widetilde{\mathbf{p}}_i = \boldsymbol{\beta}_{Li}\mathbf{p}_i \quad (7)$$

The output of the proposal-level attention is thus the $\widetilde{\mathcal{P}} = \{\widetilde{\mathbf{p}}_1, \cdots, \widetilde{\mathbf{p}}_N\}$, where N is the number of proposals.

3.2.3 Referring process

After obtaining the final representations of the two-level attentions, we then need to fuse them to localize the final target. Intuitively, given the same referring language descriptions, the model should encode the similar visual information and the language representation. Thus, we just use a simple dot-product followed by a softmax function to obtain the prediction distribution over the proposals:

$$P_i = softmax(\mathbf{h}_{t=L} \odot \widetilde{\mathbf{p}}_i) \quad (8)$$

where P_i is the probability that expression D is referring to object O_i, and $\odot$ is the dot product. Finally, we use the cross-entropy loss as the objective.

3.3. Implementation details

We optimize our model using Adam [18] in Pytorch. The learning rate is initialized to 0.001 and divided by 10 after 15 epochs with batch size of 32. The hidden state size of the LSTM is set to 512. We also set the embedding dimension for each MLP layer to 512. To avoid over-fitting, we add dropout with a ratio of 0.4 after each linear transformation in the MLP layers.

4. Experiments

To demonstrate the benefits of each module, we analyse the effect of the recurrent encoding baseline, image-level attention, proposal-level attention and the full model. Four methods are implemented and compared:

1. "**Baseline**": Compare with the full model, the baseline model doesn't use any attention mechanism. We perform mean-pooling on the extracted image feature $\mathcal{V}$ and concatenate it with the last hidden state $\mathbf{h}_L$ to obtain the image-level representation. For the proposal-level representation, we concatenate the visual representations for proposals and their corresponding spatial feature together. We further add a MLP layer to map the concatenated features to the same dimension (i.e., 512) and do dot-product to calculate the final scores over the objects.

2. "**Image-level attention**": Based on the baseline model, we further add the recurrent image-level attention as described in Sec 3.2.1.

3. "**Proposal-level attention**": Based on the baseline model, we further add the recurrent proposal-level attention as described in Sec 3.2.2, but without using the image-level attention.

Method	RefCOCO			RefCOCO+			RefCOCOg
	val	TestA	TestB	val	TestA	TestB	val
MMI [27]	-	71.72%	71.09%	-	58.42%	51.23%	62.14%
visdif [43]	-	67.57%	71.19%	-	52.44%	47.51%	59.25%
visdif+ MMI [43]	-	73.98%	76.59%	-	59.17%	55.62%	64.02%
Neg Bag [27]	-	75.60%	78.00%	-	-	-	68.40%
Luo *et al.* [25]	-	74.14%	71.46%	-	59.87%	54.35%	63.39%
Luo *et al.* (w2v) [25]	-	74.04%	73.43%	-	60.26%	55.03%	65.36%
listener [44]	77.48%	76.58%	78.94%	60.50%	61.39%	58.11%	71.12%
speaker+listener [44]	77.84%	77.50%	79.31%	60.97%	62.85%	58.58%	72.25%
speaker+listener+reinforcer [44]	78.14%	76.91%	80.10%	61.34%	63.34%	58.42%	71.72%
speaker + listener + reinforcer (ensemble) [44]	78.88%	78.01%	80.65%	61.90%	64.02%	59.19%	**72.43%**
Baseline	78.16%	77.45%	79.54%	62.41%	63.48%	59.30%	66.05%
Image-level attention	79.20%	78.29%	80.11 %	63.27 %	64.16%	60.13%	67.89%
Proposal-level attention	81.09%	80.13%	80.84%	63.57%	65.53%	60.52%	68.54%
ParalleL AttentioN (PLAN)	**81.67%**	**80.81%**	**81.32%**	**64.18%**	**66.31 %**	**61.46 %**	69.47%

Table 1: Accuracies on RefCOCO, RefCOCO+ and RefCOCOg datasets. In this setting, we use the ground-truth detection results for comparison.

Method	RefCOCO			RefCOCO+			RefCOCOg
	val	TestA	TestB	val	TestA	TestB	val
visdif [44]	-	72.03%	63.08%	-	58.87%	47.70%	58.26%
speaker+listener+reinforcer[45]	-	72.65%	62.69%	-	58.68%	48.23%	58.32%
ParalleL AttentioN (PLAN)	-	75.31%	65.52%	-	61.34%	50.86%	58.03%

Table 2: Accuracies on Refex datasets using automatic detectors.

4. "**Parallel attention**": We implement our full PLAN model based on Sec 3.

In addition, we also compare the performance of our methods against those reported in the related literature.

4.1. Datasets

We evaluate the performance on four referring expression datasets, including RefCOCO, RefCOCO+, RefCOCOg [43] and GuessWhat?! [6]. All of the datasets are collected on MS-COCO images [21].

Overall, RefCOCO has 142,210 expressions for 50,000 objects in 19,994 images. Compared to RefCOCO, RefCOCO+ removes absolute location words in referring expressions and contains 141,565 expressions for 49,856 objects in 19,992 images. Moreover, RefCOCOg has longer expressions and includes 104,560 expressions for 54,822 objects in 26,711 images. What's more, we use the "unc" standard splits for RefCOCO and RefCOCO+ datasets while using "google" split for RefCOCOg for evaluation. RefCOCO and RefCOCO+ provide person vs. object splits for evaluation. Images containing multiple people are in the 'TestA' while images containing multiple objects are in the 'TestB'. The GuessWhat?! dataset [6] is a two-player guessing game where the questioner asks a series of questions and gives answers while the guesser predicts the correct object given the questioner's information. The dataset is composed of 155,280 dialogues containing 821,889 questions/answer pairs on 66,537 unique images and 134,073 unique objects. In this paper, our method deals with the guesser task proposed in the GuessWhat dataset. Given an image and a sequence of questions and answers, the task is to predict the correct object from the set of all object candidates.

4.2. Evaluation on RefCOCO, RefCOCO+ and RefCOCOg datasets

The object candidates can be obtained through a pretrained object detection model like Faster-RCNN [33], or an object proposal model such as Edgebox [50], Objectness [1] and so on. However, for a fair comparison with previous methods, we follow [43, 44] use all the annotated entities in the image as the proposal bounding boxes at both training and test.

The results are reported in Table 1. Compared to previous state-of-the-art methods, our methods have steadily improvements on all the three datasets. Interestingly, our simple baseline model is still competitive. It proves that by calculating the similarity between the local visual features and the encoded image and language features is a reasonable framework for grounding the referring target. It can be seen that by adding the image-level attention and the proposal-level attention separately, we further observe the performance increase. This shows that recurrently attending to the informative image regions/proposals according to the language descriptions can effectively filter the noise from unrelated regions/proposals for better encoding. The effect of the proposal-level attention is more obvious.

For our full model, the two-level attention models are jointly optimized and the results are further improved. It can be attributed to that the reasonable image-level attended region is consistent with the attended proposals, which further promotes the grounding of the target object. Note that the proposed method only focus on improving the model that is similar to the listener in the latest state-of-the-art [44]. In [44], the speaker is a generative model that aims to produce referring expressions. The listener learns to embed the visual information and referring expression into a joint embedding space for comprehension and the reinforcer introduces a reward function to guide sampling of more dis-

Methods	val	test
Human	9.2%	9.2%
Random	82.9%	82.9%
LSTM [6]	37.9%	38.7%
HRED [6]	38.2%	39.0%
LSTM+VGG [6]	38.5%	39.5%
HRED+VGG [6]	38.4%	39.6%
Baseline	37.9%	39.0%
Image-level attention	37.4%	38.2%
Proposal-level attention	36.9%	37.1%
Parallel attention	**36.2%**	**36.6%**

Table 3: Classification errors for the guesser on validation and test set.

Methods	val	test
LSTM [6]	50.9%	51.4%
Baseline	48.5%	46.4%
Parallel attention	**44.3%**	**40.3%**

Table 4: Classification error rate for the guesser on validation and test set without using category features.

criminative expressions. By comparing our full model with the listener model of [44], for instance, we outperform it by nearly 5% on TestA setting of RefCOCO+. Our final single model is even better than a ensemble model in [44].

On the RefCOCOg, we outperforms the MMI [27] and other state-of-the-art on the same split. However, we observe that the performance on the RefCOCOg performs relatively worse compared with . We speculate that the length of referring expressions can have some influence on the results (e.g., RefCOCOg has length 8.5 in average while RefCOCO has an average length 3.5). Even though our improvement becomes relatively smaller, our performance still outperforms most state-of-the-arts.

4.3. Evaluation on the GuessWhat?! dataset

The results on the GuessWhat?! dataset are reported in Table 3 and Table 4. We only compare our methods with the guesser model proposed in [6] since we share the same task. In Table 3, we concatenate the object category feature with the final representation of the proposal-way framework for a fair comparison. As for the object category, we convert its one-hot class vector into a dense 512-dimensional category embedding using a learned look-up table. We can see steady improvement by comparing our approach with baseline methods in [6]. From the results, we can observe that the similar trend as shown in Sec 4.2. First, the image-level and proposal-level parallel frameworks perform as a robust baseline in grounding referential expressions. Additionally, the attention mechanism on both sides contribute a lot to the final performance increase.

During experiments, we find that the category feature has significant impact to the results. To eliminate its effect, we remove the category feature and separately report the results

in Table 4. From the table, we can see that the performance gap becomes much larger compared to Table 3, which further proves the effectiveness of the proposed methods. Especially, the LSTM model (the best model in [6]) considers the dialogue as one flat sequence to encode it into an embedding vector. By comparing our recurrent encoding baseline model with it, we observe obvious performance increase (*e.g.* 5.0% relative increase on the test set).

4.4. Recurrent parallel attention visualization

We also qualitatively illustrate the attention evolution process along with the input expression query and dialog. The results are shown in Figure 3. From the results, we can observe that the image-attention and proposal-way attention both change reasonably according to the input expression or dialog for reasoning correct target object. For example, in Figure 3, when the questioner speaks a dialog "Is it a bowl? No", the guesser is confusing with which object is the questioner referring because of the limited information. However, when the guesser model listens to more clues, we can see the image-level attention scores on the target turns to be quite high finally while the proposal-level attention also select the first proposal as the target object. It explicitly shows that the guesser model progressively feel more confident to localize the first proposal as the referring target. And we also show a failure case which corresponds to the last row in Figure 3. We can see that the attention score on the target proposal becomes smaller during steps going. Even though the image-way attention and the final result is right, the trend of proposal-way attention fails.

5. Conclusion

In this paper, we have proposed to solve the referring expression comprehension task using a novel parallel attention network. We have proposed to recurrently discover the object with variable-length language descriptions, from phrases to dialogs. To achieve this goal, we employ a two-way attention mechanism to localize the referring object on the global contextual features from the whole image and to select the referring proposal simultaneously. Since we use a recurrent style to discover the object, we make a step towards the model interpretability. With extensive experiments, we validate the advantage of our proposed methods and produce the state-of-the-art performance on several benchmarked datasets.
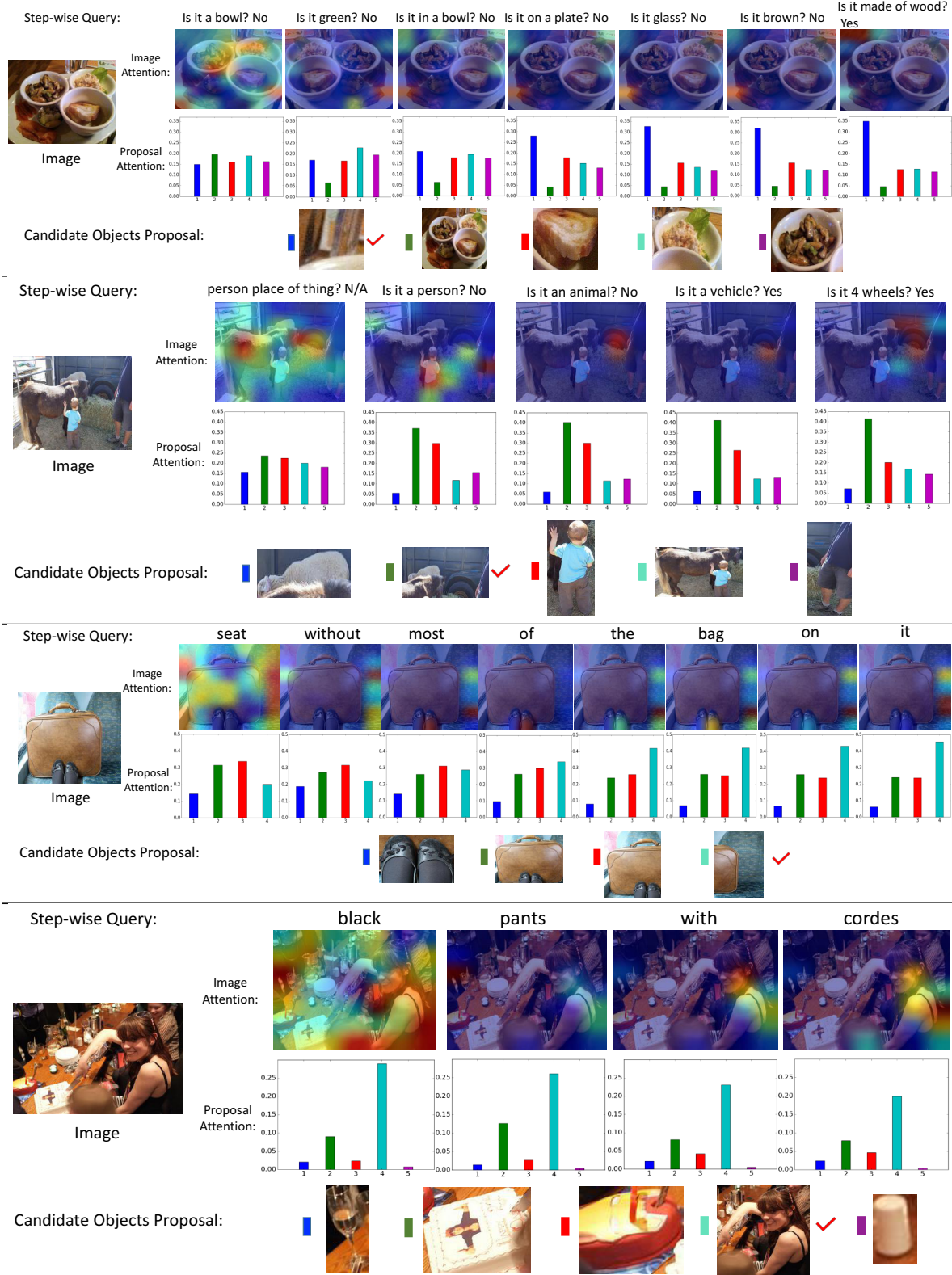
Figure 3: Qualitative results. We explicitly show the recurrent parallel attention change during inference to make the referring process explainable and visible. The referring queries are shown above the image-level attention maps. The histograms indicate the proposal-level attention on the top-5 proposals. The ground truth answer is marked. Notice the last row is a fail case.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2189–2202, 2012.

[2] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[3] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[4] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra. Visual dialog. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[5] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[6] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[7] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. Stylenet: Generating attractive visual captions with styles. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[8] C. Gan, Y. Li, H. Li, C. Sun, and B. Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[9] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *Proc. Int. Conf. Mach. Learn.*, pages 2839–2848, 2016.

[10] J. Gu, G. Wang, J. Cai, and T. Chen. An empirical study of language cnn for image captioning. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[11] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[12] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[13] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017.

[14] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4555–4564, 2016.

[15] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4565–4574, 2016.

[16] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3128–3137, 2015.

[17] A. Karpathy, A. Joulin, and F. F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1889–1897, 2014.

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. Dense-captioning events in videos. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[20] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang. Identity-aware textual-visual matching with latent co-attention. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

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

[22] H. Lu, L. Zhang, Z. Cao, W. Wei, K. Xian, C. Shen, and A. van den Hengel. When unsupervised domain adaptation meets tensor representations. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[23] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017.

[24] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 289–297, 2016.

[25] R. Luo and G. Shakhnarovich. Comprehension-guided referring expressions. *arXiv preprint arXiv:1701.03439*, 2017.

[26] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 11–20, 2016.

[27] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *Proc. Eur. Conf. Comp. Vis.*, pages 792–807, 2016.

[28] N. Patricia and B. Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1442–1449, 2014.

[29] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[30] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2641–2649, 2015.

[31] R. Qiao, L. Liu, C. Shen, and A. van den Hengel. Less is more: zero-shot learning from online textual documents with noise suppression. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2249–2257, 2016.

[32] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proc. Int. Conf. Mach. Learn.*, pages 759–766, 2007.

[33] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In

International Conference on Neural Information Processing Systems, pages 91–99, 2015.

[34] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *Proc. Int. Conf. Mach. Learn.*, pages 1842–1850, 2016.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Repren.*, 2015.

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3156–3164, 2015.

[37] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng. Structured matching for phrase localization. In *Proc. Eur. Conf. Comp. Vis.*, pages 696–711. Springer, 2016.

[38] 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. Comp. Vis. Patt. Recogn.*, pages 203–212, 2016.

[39] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4622–4630, 2016.

[40] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. Int. Conf. Mach. Learn.*, pages 2048–2057, 2015.

[41] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov. Review networks for caption generation. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2361–2369, 2016.

[42] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4651–4659, 2016.

[43] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *Proc. Eur. Conf. Comp. Vis.*, pages 69–85, 2016.

[44] L. Yu, H. Tan, M. Bansal, and T. L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[45] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[46] L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[47] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma. Structured attentions for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[48] B. Zhuang, L. Liu, C. Shen, and I. Reid. Towards context-aware interaction recognition for visual relationship detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[49] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. v. d. Hengel. Hcvrd: a benchmark for large-scale human-centered visual relationship detection. In *Proc. AAAI Conf. on Arti. Intel.*, 2018.

[50] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.