Language-Conditioned Region Proposal and Retrieval Network for Referring Expression Comprehension

Yanwei Xie
University of Science and Technology of China
HeFei, China
yanweiya@mail.ustc.edu.cn

Daqing Liu
University of Science and Technology of China
HeFei, China
liudq@mail.ustc.edu.cn

Xuejin Chen
University of Science and Technology of China
HeFei, China
xjchen99@ustc.edu.cn

Zheng-Jun Zha
University of Science and Technology of China
HeFei, China
zhazj@ustc.edu.cn

ABSTRACT

Referring expression comprehension (REC) is a multi-modal task that aims to localize target regions in images according to language descriptions. Existing methods can be concluded into two categories, proposal-based methods and proposal-free methods. Proposal-based methods first detect *all* candidate objects in the image and then retrieve the target among those objects based on the language description, while proposal-free methods directly locate the region based on the language without *any* region proposals. However, the proposal-based methods suffer from separate region proposal networks that actually do not suit this task well, and the proposal-free methods are not able to perform fine-grained visual-language alignments to yield higher precision. To overcome the above drawbacks, we propose a language-conditioned region proposal and retrieval network that first detects those regions only related to the language and then retrieves the target region by compositional reasoning on the language. Specifically, the proposed network consists of a language-conditioned region proposal network (LC-RPN) to detect those language-related regions, and a language-conditioned region retrieval network (LC-RRN) to perform region retrieval with a full understanding of the language. A pre-training mechanism is proposed to teach our model knowledge about language decomposing and vision-language alignment. Experimental results demonstrate that our proposed method achieves leading performance with high inference speed on RefCOCO, RefCOCO+, and RefCOCOg benchmarks.

CCS CONCEPTS

• **Computing methodologies** → *Lexical semantics*; **Information extraction**; *Scene understanding*; **Matching**; **Object identification**.

MMPT '21, August 21, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8530-5/21/08…$15.00
https://doi.org/10.1145/3463945.3469055

KEYWORDS

Region proposal; Multi-modal retrieval; Referring expression comprehension; Multi-modal pre-training

ACM Reference Format:
Yanwei Xie, Daqing Liu, Xuejin Chen, and Zheng-Jun Zha. 2021. Language-Conditioned Region Proposal and Retrieval Network for Referring Expression Comprehension. In *Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding (MMPT '21), August 21, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3463945.3469055

1 INTRODUCTION

Referring expression comprehension (REC) as a relatively new task attracts increasing attention in recent years for it is the foundation of many high-level multi-modal applications, *e.g.*, image captioning [16, 32, 35], video captioning [33], visual question answering [1], vision-language navigation [8] and video moment localization [23]. The REC task takes an image and a sentence that unambiguously describes an object in the image as input, and output a bounding box location that localizes the described object. Its key challenge lies in how to predict the accurate location of objects and how to distinguish the target object from the context objects according to language descriptions.

Current approaches for REC can be roughly divided into two categories according to the requirement of pre-detected proposals. As illustrated in Figure 1 (a), the first category is proposal-based, where those methods [17, 18, 25, 29–31] need first employ a pre-trained generic region proposal network (RPN), *e.g.*, Faster-RCNN [22], to detect *all* objects that appeared in the image, and then retrieve the target object among those objects with the language. During the retrieval phase, many recent works [17, 29] introduce compositional reasoning on the language and achieve promising performance with a good understanding of the language structure and semantic meanings. However, those methods are still facing some problems. They perform heavy computation at each candidate region, which enables them to fully make use of the referring expression, but suffer from the fully visual-based RPN which generates large amount of candidate regions, among which are pretty much non-object regions and expression-irrelevant regions. These noisy candidate regions not only introduce massive useless computations into proposal-based methods, but also impede the learning procedure of following modules. For example, those methods will consider the relationship

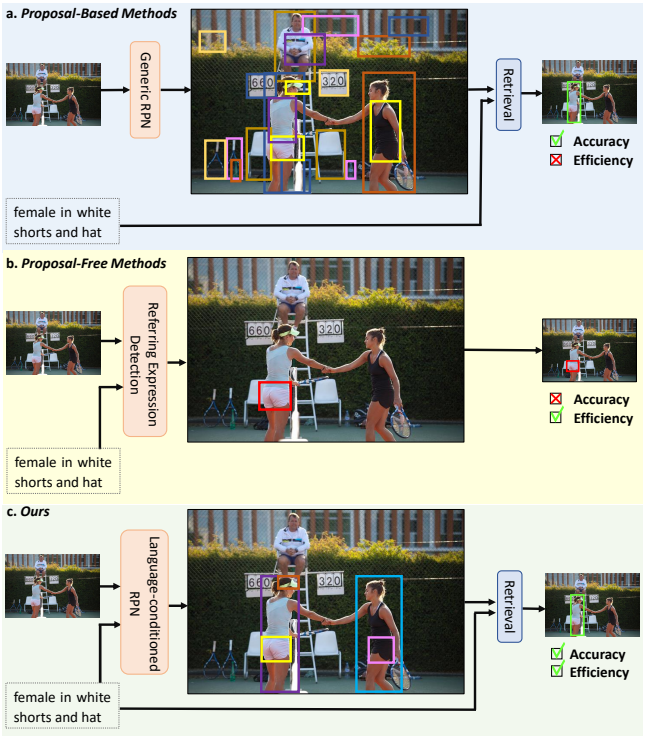

Figure 1: The illustration of three different frameworks for REC. (a) The proposal-based methods first detect *all* objects in images and then retrieve the target. (b) The proposal-free methods directly predict the location of the target without any region proposals. (c) Our Language-conditioned Region Proposal and Retrieval Network first detects several objects related to the language, and then performs retrieval to produce the final results.

between the chair and the scoreboard (as is shown in Figure 1) which are never mentioned in the referring expression.

In contrast, as illustrated in Figure 1 (b), the second category is proposal-free, where those methods [2, 13, 26, 27] directly incorporate the referring expression detection process into some one-stage object detection frameworks, *e.g.*, YOLO [21] and CenterNet [5], rather than employ a separate generic RPN. Those methods greatly increase the inference speed but they also introduce new issues. Since they usually represent the language as a whole feature [13, 27] or design a multi-round processing [26], the linguistic structure and semantic meaning of the referring expressions are usually neglected. Therefore, the accuracy of those methods is limited because they can hardly distinguish the target object from context objects, especially those objects of the same class (*e.g.*, the left female and the right female in Figure 1) and also may misunderstand the target subject noun (*e.g.*, the "female" and the "white shorts").

In this paper, we propose a novel framework for referring expression comprehension, named as Language-Conditioned Region Proposal and Retrieval Network, to solve the above disadvantages of both proposal-based and proposal-free methods. As illustrated in Figure 1 (c), the proposed method first detect those object regions related to the referring expression (*e.g.*, female, shorts, hat) with a Language-Conditioned Region Proposal Network (LC-RPN), and then perform compositional reasoning on the language and retrieve the localization result with a Language-Conditioned Region Retrieval Network (LC-RRN). Specifically, the LC-RPN is based on the CenterNet and we further improve the CenterNet by taking the language representation as the filter kernels inspired by RCCF [13]. However, different from RCCF, our LC-RPN is pre-trained on Visual Genome [12] to enable it to produce multi-region proposals. The LC-RRN first decomposes the language into four parts, *i.e.*, subject, attribute, relationship, and unimportant, by the Language Attention Module. Further, the LC-RRN contains three reasoning modules, as Subject Module, Attribute Module and Relationship Module, to perform compositional reasoning. After that, we weighted sum each module output as the retrieval criterion to produce the final result. The training process of the proposed method consists of the pre-training stage on Visual Genome and the fine-tuning stage on the REC datasets.

With the proposed method, we break the bottleneck of efficiency by reducing the number of candidate regions and release the limitation of the accuracy by performing compositional reasoning to retrieve the target object. We validate the effectiveness of the proposed method through extensive experiments on the RefCOCO, RefCOCO+, and RefCOCOg datasets. The proposed method achieves very competitive performances against the state-of-the-art methods of both proposal-based and proposal-free. Meanwhile, the inference speed of the proposed method reaches 20FPS (49ms) with a single GTX1080Ti. We also show promising qualitative results of language attention over the whole sentences.

2 RELATED WORK

2.1 Proposal-based REC

Proposal-based methods model the task of referring expression comprehension into a detect-and-retrieve pipeline. They first adopt off the shelf tools (*e.g.*, [22, 36]) to generate hundreds of ROIs, and then compute match scores between referring expression and each ROI. [10, 19, 30] force the visual and language embedders to learn more general knowledge by jointly generating and comprehending the referring expression. [9, 29] learn to decompose the expression into different components and match each component with the corresponding visual region through a modular network. [18, 24, 25] build graphs of the objects in the image, where the nodes and edges correspond to the objects and inter-object interactions. With visual and language attention mechanism, relationships between objects can be recognized and help to reason in the graph to determine the region that best matches the expression. [3, 17] parse input expression with external language parser, and apply different modules to different syntactic components and accumulate confidence for each proposed region. The most related proposal-based method to us is [29], which learns to generate three phrase-level representations emphasizing information about subject, location, and relationship respectively from referring expression. However, phrase-level representations contain noise from inessential words. Our model learns to automatically identify the subject word, attribute words, and

relationship words in expressions, and only pass these key words to the retrieval module.

2.2 Proposal-free REC

Proposal-free REC methods model this task into a language-based object detection pipeline. [27] concatenates the language feature of referring expression with the image feature map, and fed them into a fusion module followed by a prediction head to predict the bounding box of the referent. [13] filter image feature maps with language kernels to predict the center point of the referred region, and regress bounding box at each location. [13] adopt a Deep Layer Aggregation network[28] to obtain feature pyramid for the image, and fuse these feature maps with deformable convolution[4]. On the other hand, referring expression is fed into a bi-LSTM to obtain a language representation, which is later transformed into three 64-D language kernels to perform correlation filtering on image feature maps. Proposal-free models seldom refer to linguistic priori information during inference, in other words, they did not benefit from human knowledge. In this paper, we propose a region proposal network following the proposal-free REC pipeline to generate the expression-related object, and employ linguistic priori information to re-rank the object proposals.

3 METHOD

In this section, we introduce our new model for fast and accurate referring expression comprehension, which we call **L**anguage-**C**onditioned **R**egion **P**roposal and **R**etrieval Network (LCRPR). LCRPR is conceptually simple: proposal-free REC models (e.g. [13, 27]) fuse language feature and visual feature in a straightforward way with an absence of linguistic priori information, thus leave space for improvement. The overview of LCRPR is illustrated in Figure 2. As is shown in Figure 2, LCRPR consists of two parts, which are Language-Conditioned Region Proposal Network (LC-RPN) and Language-Conditional Region Retrieval Network (LC-RRN). The LC-RPN detects expression-related objects in the image, and the LC-RRN uses a modular attention network to compute match scores for each detected object. The entire model is trained end-to-end, thus the LC-RPN and LC-RRN mutually guides each other to care for vital information in referring expression.

3.1 LC-RPN

State-of-the-art proposal-free REC models are yielding higher precision and efficiency than state-of-the-art proposal-based methods, because proposal-free models do not suffer from the misalignments between generated proposals and the ground-truth objects, and they do not perform various complex operations to a large number of region proposals. We propose to generate a small number of accurate object proposals through the proposal-free referring expression comprehension pipeline, thus help to yield higher performance for proposal-based REC methods.

3.1.1 Language representation. To get language representation of referring expression, we first embed each word in the expression into 1024-D vector, followed by a fully connected layer to transform the vector into 512-D. Then, a bi-LSTM is adopted to get representation of each word with language contexts encoded, which are the concatenation of the hidden vectors in both directions, denoted as:

$$L_w = \left\{ \left[\overrightarrow{h}_t, \overleftarrow{h}_{T+1-t} \right] \mid t \in [1, T] \right\} \tag{1}$$

Besides, we represent the whole expression with the last hidden state of both directions: $L_E = \left[\overrightarrow{h}_T, \overleftarrow{h}_T \right]$, which is later used to generate module weights and to detect expression-related objects.

3.1.2 Cross-modality Correlation Filtering Region Proposal Network. We follow [13] to build a region proposal network that takes both image and expression as input. We adopt three fully connected layers to generate three filtering kernels from expression representation L_E, and further generate three heatmaps through correlation filtering, which are later averaged to predict the location of objects that appeared in the expression. On the other hand, a regression network is adopted to regress bounding box at each location. ROI Align[6] followed by a 3-layer residual network[7] is adopted to extract object features from last visual feature map. We denote the object features as $\{v_i\}$.

3.2 LC-RRN

3.2.1 Language attention module. Different linguistic components in referring expressions have inherently different patterns for cross-modality alignments. Consider the following two referring expressions that refer to the same region in an image:

- *The red balloon*
- *The balloon held by the girl*

It is natural and reasonable to attend to appearance for first expression, and attend to inter-object relationship for second expression. We propose a language attention module to learn to identify subject word, attribute words and relationship words in referring expression. We call these words as *keywords*. To be exact, the language attention module feeds L_w into a fully connected layer to get T 4-D vectors, and adopt softmax to generate T 4-D distributions:

$$\left[p^t_{\text{subj}}, p^t_{\text{attr}}, p^t_{\text{rela}}, p^t_{\text{unimportant}} \right] = \text{softmax}\left(\text{MLP}\left(L^t_w \right) \right) \tag{2}$$

Each word is classified to *subject word*, *attribute word*, *relationship word* or *unimportant word*, according to theirs predict 4-D distributions, and the corresponding probability is their classification confidence. Then, we adopt three different modules to compute match score between each expression-related object and each keyword. In addition, language attention module computes three weights from expression representation L_E to assign weights to the modular matching scores:

$$\left[w_{\text{subj}}, w_{\text{attr}}, w_{\text{rela}} \right] = \text{softmax}\left(\text{MLP}\left(L_E \right) \right) \tag{3}$$

3.2.2 Attribute module. We align candidate objects and attribute words with a matching function, which uses two MLPs to transform object feature and word representation into a common embedding space, and use the sigmoid activation of the inner-product of two embeddings to measure the match score between candidate object and attribute word:

$$S_{attr}\left(o_i \mid w_t \right) = \text{sigmoid}\left(MLP\left(v_i \right) \cdot MLP\left(L^t_w \right) \right) \tag{4}$$

This match function is used in subject module and relationship module as well. The overall attribute score for each object is the average of its match scores to all attribute words in expression.

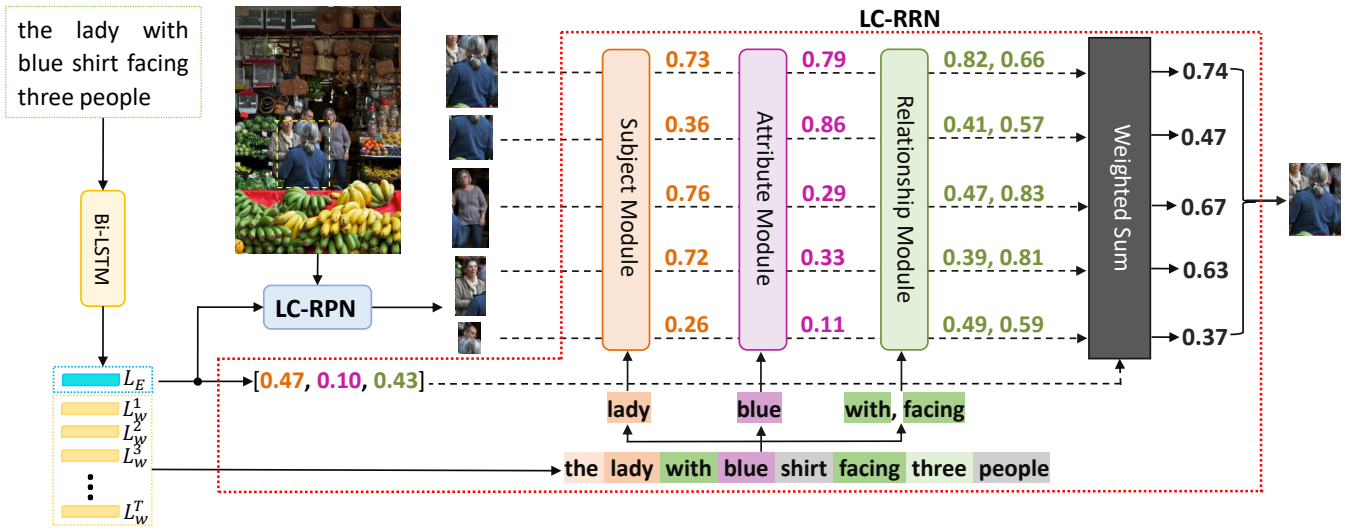

Figure 2: The overview of the proposed LCRPR framework. (a) LC-RPN: LC-RPN takes the image and expression as input, and detects expression-related objects. The LC-RPN shares the visual feature map with LC-RRN. (b) LC-RRN: LC-RRN adopts a language attention module to identify the subject (orange) word, attribute (purple) words, and relationship (green) words in expression, which we call *keywords*, and adopt three different modules to compute match scores between each keyword and each detected object. The overall match score of each object is the weighted sum of three module scores, where the module weights are computed from expression representation.

3.2.3 Subject module. Subject word is the word in expression that names the referent, it is an ensemble of all inherent or abstract attributes of the referent. For example, in description *"the red balloon held by a girl"*, *"balloon"* is the subject word, meanwhile *"red"* is an attribute word describing specific attribute of the subject. Therefore, we treat subject word in the same way as attribute words, the only difference is subject word only describes the referent, while attribute words could describe the referent and any context object. Subject module have identical structure with attribute module, but with different parameters. This module computes match scores between the subject word and each candidate object:

$$S_{subj}\left(o_i \mid w_t\right) = \text{sigmoid}\left(MLP\left(v_i\right) \cdot MLP\left(L_w^t\right)\right) \tag{5}$$

3.2.4 Relationship module. Similar to [29], for each candidate object and each relationship word, we first compute the offset of other objects to the candidate object, and chose up to five objects whose offset is relatively small as supporting objects. Then, we concatenate the feature of each supporting object with its offset to the candidate object and feed it into a MLP to get a 64-D vector. On the other hand, embedding of the relationship word is transformed into 64-D with another MLP. With inner-product and sigmoid, we get the score of each candidate-object pair:

$$\begin{aligned} S_{\text{rela}}\left(\left(o_i, o_j\right) \mid w_t\right) &= \text{sigmoid}\left(MLP\left(x_i^j\right) \cdot MLP\left(L_w^t\right)\right), \\ x_i^j &= MLP\left(\left[v_j, p_j^i\right]\right), \end{aligned} \tag{6}$$

Where x_i^j is the representation of supporting object, and p_j^i is the supporting object's offset[29] to the candidate object. We chose the highest score as the match score between the candidate object and the relationship word, and take the average of the candidate object's match score to all relationship words in expression as its relationship score. The procedure is illustrated in Figure 3.

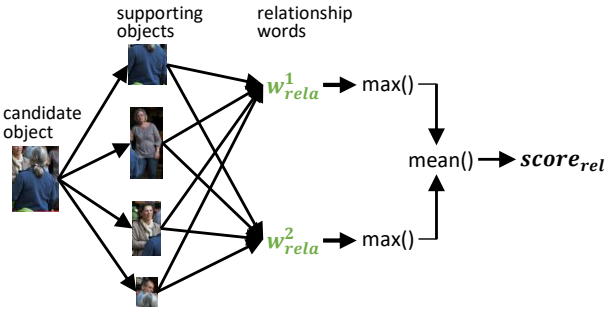

Figure 3: The procedure of the relationship module.

3.3 Vision-Language Pre-training

3.3.1 Dataset introduction. Visual Genome[12] is a large-scale visual-language dataset containing 108,077 images and millions of corresponding visual and language labels. To be exact, for each image in Visual Genome dataset, there are averagely: 50.04 region-description pairs, 35.18 annotated objects with bounding-box, name and attribute labels, 21.43 annotated inter-object relationships with predicate words, and so on. Different from RefCOCO series datasets, region-description pairs in Visual Genome dataset are not annotated based on annotated objects, thus have independently annotated bounding-boxes, and some regions do not indicate an unique meaningful visual entity. Moreover, many region descriptions in

Visual Genome dataset cannot unambiguously refer to one single region. Therefore, region-description pairs in Visual Genome dataset are not suitable for the training and evaluation of REC models. Nevertheless, we believe with appropriate data preprocessing, comprehensive vision-language annotations in Visual Genome dataset can be applied to pretrain our REC model, thus teach our model some general knowledge about phrase decomposing and vision-language alignment. Next, we will detail the preprocess of Visual Genome dataset.

3.3.2 Dataset preprocessing. For every region-description pair, we first parse the phrase with an external NLP parser[1] to get part-of-speech tags for each word, and then identify all noun words. Second, for each noun word in the phrase and each annotated object in the image, we calculate the semantic similarity[2] between each noun word and each object's name. Then, objects having high semantic similarity with any of the noun words in the phrase are considered as description-related objects. Third, as some region descriptions in Visual Genome do not refer to certain object (e.g. *"the area is blue"*), we delete region-description pairs whose region do not intersect with any of its description-related objects (including those who have no description-related object, and here we define *intersect* as intersection-over-union > 0.5), and take the description-related object that have highest intersection-over-union with the described region as the subject. For words in the phrase, we take the one having highest semantic similarity with the subject's name as ground-truth subject word to form phrase-subject pair. Besides, for ground-truth attribute-object pairs in the image, if the attribute word is in the description phrase and the object is description-related, we take the attribute-object pair to supervise attribute module and take the attribute word to form phrase-attribute pair to supervise the language attention module. For ground-truth inter-object relationship pairs in image, if both the subject and the object are description-related and the predicate is in the description phrase, we take the inter-object relationship pair to supervise relationship module and take the predicate to form phrase-relationship pair to supervise the language attention module. In this way, we get a dataset containing 1278056 samples, each sample has a region-description pair, a phrase-subject pair and at least one description-related object. Besides, among all the samples, 624757 samples have at least one attribute-object pair and phrase-attribute pair, and 267352 samples have at least one inter-object relationship pair and phrase-relationship pair. We illustrate a sample from our preprocessed Visual Genome dataset in Figure 4, which contains one attribute-object pair and two inter-object relationship pairs.

3.3.3 Pre-training loss. Following [13], we adopt a gaussian kernel to splat the center points of ground-truth bounding boxes of described region and ground-truth description-related objects in a heatmap, and compute focal loss[14] L_c to supervise the center points prediction procedure, where the center point of the described region gets slightly higher weight than center points of description-related objects. Besides, L1 loss function is adopted to regress the bounding boxes, we denote this loss as L_{reg}.

[1]http://www.nltk.org/install.html
[2]http://www.nltk.org/howto/wordnet.html

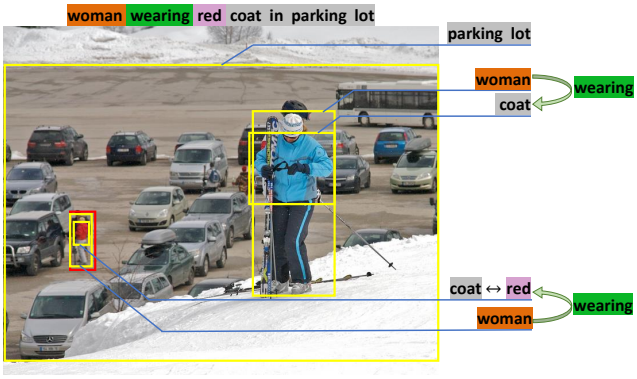

Figure 4: A sample in our preprocessed Visual Genome dataset. The red box shows the ground-truth region, and yellow boxes are ground-truth description-related objects.

During pre-training, language attention module is supervised by ground-truth phrase-subject pairs, phrase-attribute pairs and phrase-relationship pairs. The loss function is defined as:

$$\mathrm{L_m} = -\frac{1}{\mathrm{sum}\left(\mathrm{is_m}\right)}\sum_{\mathrm{i}=1}^{\mathrm{T}}\mathrm{is_m^i}\log\left(\mathrm{p_m^i}\right) - \frac{1}{\mathrm{sum}\left(\mathrm{not_m}\right)}\sum_{\mathrm{i}=1}^{\mathrm{T}}\mathrm{not_m^i}\log\left(1-\mathrm{p_m^i}\right) \tag{7}$$

$$L_{\text{attention}} = L_{\text{subject}} + L_{\text{attribute}} + L_{\text{relationship}} \tag{8}$$

where $\mathrm{m} \in \{$ subject, attribute, relationship $\}$, and $\mathrm{p_m^i}$ is the predicted likelihood of the i-th word being a m-word. $\mathrm{is_m^i} = 1$ and $\mathrm{not_m^i} = 0$ when the i-th word in phrase is a ground-truth m-word, otherwise $\mathrm{is_m^i} = 0$ and $\mathrm{not_m^i} = 1$. When there is no ground-truth m-word for current sample, $\mathrm{L_m} = 0$. Hinge loss is adopted to supervise the learning procedure of subject module, attribute module and relationship module. In subject module, the ground-truth pair is denoted as $\left(\mathrm{w_{subj}}, \mathrm{o_{subj}}\right)$, we randomly sample an object of different category with the subject to form a negative pair $\left(\mathrm{w_{subj}}, \mathrm{o_{context}}\right)$. We force the match score of ground-truth pair to be greater than negative pair with a margin α_1. In attribute module, for each ground-truth attribute-object pair (a_i, o_i), we randomly sample two negative pairs (a_j, o_i) and (a_i, o_j), we force the match score of ground-truth attribute-object pairs to be greater than match scores of negative attribute-object pairs with a margin α_2. Only samples with ground-truth attribute-object pairs go through the attribute module. In relationship module, for each ground-truth inter-object relationship pair $(subj_i, predicate_i, obj_i)$, we take the reversed pair $(obj_i, predicate_i, subj_i)$ as negative sample, and force the match score of ground-truth inter-object relationship pair to be greater than match score of negative pair with a margin α_3. Only samples with ground-truth inter-object relationship pairs go through the relationship module.

$$L_{\text{subj}} = \max\left(0, \alpha_1 + S_{\text{subj}}\left(\mathrm{o_{context}} \mid \mathrm{w_{subj}}\right) - S_{\text{subj}}\left(\mathrm{o_{subj}} \mid \mathrm{w_{subj}}\right)\right) \tag{9}$$

$$L_{\text{attr}} = \begin{cases} \frac{1}{N_{\text{attr}}} \sum_{i=1}^{N_{attr}} \Big[\max\left(0, \alpha_2 + S_{\text{attr}}\left(o_i \mid a_j\right) - S_{\text{attr}}\left(o_i \mid a_i\right)\right) + \\ \max\left(0, \alpha_2 + S_{\text{attr}}\left(o_j \mid a_i\right) - S_{\text{attr}}\left(o_i \mid a_i\right)\right) \Big], N_{\text{attr}} > 0 \\ \\ 0, N_{\text{attr}} = 0 \end{cases} \tag{10}$$

$$L_{\text{rela}} = \begin{cases} \frac{1}{N_{\text{rela}}} \sum_{i=1}^{N_{\text{rela}}} \Big[\max\left(0, \alpha_3 + S_{\text{rela}}\left(\left(\text{obj}_i, \text{subj}_i\right) \mid \text{predicate}_i\right) - \\ S_{\text{rela}}\left(\left(\text{subj}_i, \text{obj}_i\right) \mid \text{predicate}_i\right) \Big], N_{\text{rela}} > 0 \\ \\ 0, N_{\text{rela}} = 0 \end{cases} \tag{11}$$

The overall loss at pre-training stage incorporates the center points prediction loss, the bounding box regression loss, the attention loss and three modular losses:

$$L_{\text{pretrain}} = L_c + L_{\text{reg}} + \lambda_1 L_{\text{attention}} + \lambda_2 \left(L_{\text{subj}} + L_{\text{attr}} + L_{\text{rela}}\right). \tag{12}$$

3.4 Fine-tuning

Our model is separately fine-tuned and evaluated on RefCOCO[11], RefCOCO+[11] and RefCOCOg[19] datasets. LC-RPN takes the ground-truth referent and ground-truth objects near ground-truth referent as supervision.

During fine-tuning, language attention module is not directly supervised, it predicts subject word, attribute words and relationship words, which are later feed into three modules to compute modular matching scores for each expression-related object detected by LC-RPN. Following [29], three module weights are computed from the expression representations L_E, and the overall match score of each expression-related object is the weighted sum of three module scores. Hinge loss is adopted here to compute the rank loss:

$$L_{\text{rank}} = \max(0, \alpha + S(\tilde{r} \mid e) - S(r \mid e)) + \max(0, \alpha + S(r \mid \tilde{e}) - S(r \mid e)) \tag{13}$$

where e is the referring expression, r is the ground-truth region, $\tilde{e}$ is a randomly sampled expression that refers to a different region in the same image, and $\tilde{r}$ is a randomly sampled region in the image whose intersection-over-union with the ground-truth region is less than 0.25. Finally, the overall loss during fine-tuning is:

$$L = L_c + L_{reg} + \lambda L_{\text{rank}} \tag{14}$$

During evaluation, we pick the region with highest match score as the prediction.

4 EXPERIMENTS

4.1 Experimental setting

4.1.1 Dataset. We conduct experiments on three popular datasets, which are RefCOCO, RefCOCO+ and RefCOCOg. These three datasets are all built on the subsets of Microsoft COCO images[15]. RefCOCO and RefCOCO+ use the same image subset of Microsoft COCO dataset, containing 3.9 same-type objects on average in each image. Referring expressions in RefCOCO and RefCOCO+ datasets tend to be short phrases with an average length of 3.5 words. Besides, RefCOCO+ dataset forbids using absolute location words in referring expressions, making the expression more focused on appearance and inter-object relationships. RefCOCOg has longer expressions with averagely 8.4 words. Besides, images in RefCOCOg contain only 1.63 same-type objects on average. Both RefCOCO and RefCOCO+ contain approximately 142k referring expressions for 50k referents in 20k images, and due to the large number of same-type objects in images, these two datasets are split into 4 sets, which are train, validation, testA and testB. TestA contains images with multiple people and testB contains images with multiple objects. RefCOCOg contains 95010 referring expressions for 49822 objects in 25799 images, and has two types of partitions. The first[19] divides the dataset by randomly partitioning objects into training and validation sets and the same image could appear in both training and validation set. The second partition[20] is composed by randomly partitioning images into training, validation and testing sets, we run experiments on this split.

4.1.2 Evaluation Metric. We follow previous works and evaluate our model with Prec@0.5 metric. To be exact, for each referring expression, the predicted bounding box is correct if its intersection-over-union with the ground-truth bounding-box is greater than 0.5. The model's precision is evaluated by the ratio of correctly grounded referring expressions to the general number of referring expressions in test sets.

4.2 Implementation details

During vision-language pre-training and fine-tuning, three rank modules take ground-truth objects as input. During test, we pass 10 top scored region proposals to the rank modules. During fine-tuning and test, words in expressions are classified as subject word, attribute word, relationship word or unimportant word, according to their predicted 4-D distributions. A good rule of thumb during fine-tuning and test is that when more than 2 words in an expression are classified as attribute word or relationship word, we only pass the two with highest classification confidence to the corresponding rank module. While for subject module, only the subject word with highest classification confidence is passed to subject module during fine-tuning and test.

Following [13], the visual backbone is pretrained with object detection task[34] on MS-COCO dataset, with images in validation sets and test sets of RefCOCO, RefCOCO+ and RefCOCOg datasets removed. The vision-language pre-training is conducted on our preprocessed Visual Genome dataset for 50 epochs, each epoch uses 500000 samples. The learning rate is 5e-4, and drop to 5e-5 after the 40th epoch.

Our model is separately trained and evaluated at RefCOCO, RefCOCO+ and RefCOCOg datasets. We train our model for 80 epochs. The initial learning rate is 2.5e-4, which decreases 10 folds at the 60th epoch, and again at the 70th epoch.

4.3 Comparison to the State-of-the-art

We compare our model with state-of-the-art methods in RefCOCO, RefCOCO+ and RefCOCOg dataset, the results are shown in table 1. we surpassed [13] in all three datasets with a big margin, this is due to the priori-linguistic-information-guided reasoning mechanism. On the other hand, our model achieved higher precision and faster speed compared to [29], we attribute this improvement to our new region proposal network, which takes referring expression into account and generates small amount of accurate region proposals. Our model takes approximately 69% more time than [13] for each

			RefCOCO			RefCOCO+			RefCOCOg		
	Method	Visual Encoder	val	testA	testB	val	testA	testB	val	test	Time (ms)
proposal based	S+L+R[31]	vgg16	-	73.78	63.83	-	60.48	49.36	-	59.84	-
	DGA[25]	vgg16	-	78.42	65.53	-	69.07	51.99	-	63.28	341
	MattNet[29]	res101-mrcn	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27	320
	NMTree[17]	res101-frcn	76.41	81.21	70.09	66.46	72.02	57.52	65.87	66.44	-
proposal free	SSG[2]	DarkNet-53	-	76.51	67.5	-	62.14	49.27	58.8	-	25
	FAOA[27]	DarkNet-53	72.05	74.81	67.59	55.72	60.37	48.54	59.03	58.7	23
	ReSC[26]	DarkNet-53	77.63	80.45	72.3	63.59	68.36	56.81	67.3	67.2	36
	RCCF[13]	DLA-34	-	81.06	71.85	-	70.35	56.32	-	65.73	29
ours	LCRPR	DLA-34	**80.26**	**83.52**	**74.97**	**66.86**	**72.83**	**59.17**	**69.75**	**69.48**	49

Table 1: Comparison to state-of-the-art approaches on RefCOCO, RefCOCO+ and RefCOCOg.

		Refcoco			Refcoco+			Refcocog	
		val	testA	testB	val	testA	testB	val	test
1	MattNet[29]	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
2	RCCF[13]	-	81.06	71.85	-	70.35	56.32	-	65.73
3	LC-RPN	72.16	76.21	67.19	59.73	66.54	52.56	63.49	63.40
4	LC-RPN + subj	75.27	79.86	70.07	63.24	69.64	55.43	65.38	65.21
5	LC-RPN + subj + rela	77.84	81.12	72.67	64.88	70.76	57.35	66.71	66.64
6	LC-RPN + subj + attr	78.27	81.57	72.94	65.12	70.94	57.73	67.28	67.34
7	LC-RPN + subj + attr + rela	79.08	82.41	73.83	65.91	71.62	58.45	68.64	68.29
8	LC-RPN + subj + attr + rela + VLpretrain	**80.26**	**83.52**	**74.97**	**66.86**	**72.83**	**59.17**	**69.75**	**69.48**

Table 2: Ablation experiments on RefCOCO, RefCOCO+ and RefCOCOg.

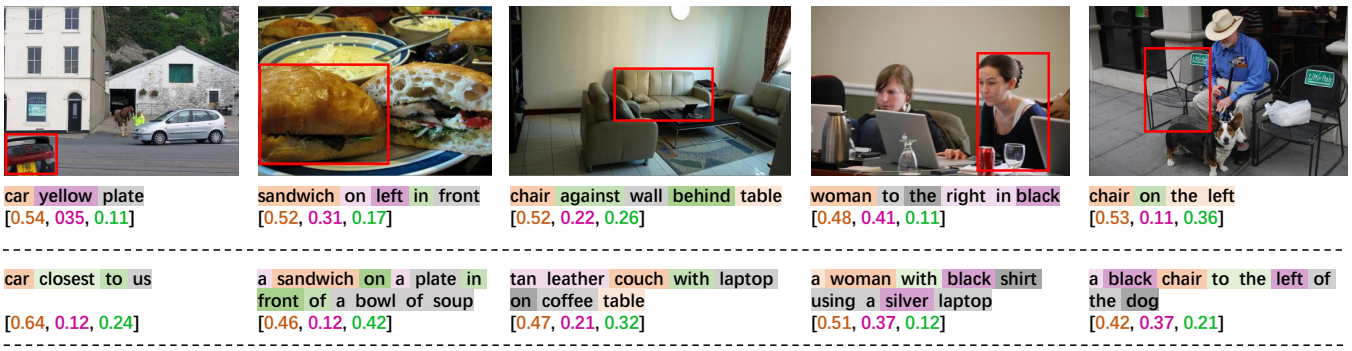

Figure 5: Visualization of language attention. For each region, we randomly picked two expressions from validation or test sets of RefCOCO, RefCOCO+ or RefCOCOg. Words predicted as subject word, attribute word, relationship word and unimportant word are typed with orange, purple, green and grey bases, respectively. Orange, purple and green numbers are computed weights for subject module, attribute module and relationship module, respectively.

region-expression pair during inference. However, considering the significant precision improvement, we believe this is basically a good tradeoff. Nevertheless, our model is still much faster than traditional proposal-based methods.

4.4 Ablation studies

To better understand the importance of each module, we conducted abundant ablation experiments on three datasets. The results are shown in table 2.

4.4.1 LC-RPN. We individually trained our LC-RPN with ground-truth region and ground-truth objects, without vision-language pre-training. During test, we chose the region proposal with highest score as the predicted referent, the results are shown in row 3. We observed a significant decline compared to [13]. However, the results are still not too bad, which largely because we attached more importance to the ground-truth referent region than context object regions during the training of LC-RPN, this setting is applied to all of our experiments in this paper.

4.4.2 Subject module. In row 4, we removed the attribute module and relationship module from full model, then we trained and tested this ablated model with the same setting as row 3 and row 7. Experiments in row 5 and row 6 followed this routine. In this experiment, we can see that subject module brings significant improvement compared to row 3, but still yields worse performance than [13, 29].

4.4.3 Attribute module and relationship module. Results in row 5 , 6 and 7 show that attribute module and relationship module are both important, and any of them can help to yield better performance than [13, 29].

4.4.4 Vision-language pre-training. Results in row 8 show that vision-language pre-training brings approximately 1 point improvement in each dataset. This is because ground-truth mappings in Visual Genome dataset encourage our model to explore more general rules about decomposing expressions and aligning regions to words.

4.5 Visualization of the language attention

Figure 5 shows some examples of language attention produced by the models in row 8 of table 2. We randomly picked 5 regions from three RefCOCO series datasets, and visualized the language attention of their referring expressions. The top row shows the regions with red box in images, and the rest rows show the automatically computed language attention and module weights corresponding to their referring expressions randomly picked from three datasets. We conclude that our language attention module is able to allocate most words with right labels, and adaptively assign appropriate weights to three rank modules. For example, our language attention module assigns small weight to relationship module when the referring expression is *"car yellow plate"*, and assigns small weight to attribute module when the referring expression is *"car closest to us"*, as is shown in Figure 5.

5 CONCLUSION

In this paper, we proposed a language-conditioned region proposal and retrieval network for fast and accurate referring expression comprehension. Different from either proposal-based or proposal-free methods, the proposed method first detects several region proposals that are related to the referring expression by a language-conditioned region proposal network, and then performs compositional reasoning on the language to retrieve the target object by a language-conditioned region retrieval network. The proposed method successfully overcomes the disadvantage of both proposal-based and proposal-free methods, and achieves high accuracy with high computing efficiency.

ACKNOWLEDGEMENT

This work was supported by the National Key RD Program of China under Grand 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants U19B2038, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025.

REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4971–4980.

[2] Xinpeng Chen, Lin Ma, Jingyuan Chen, Zequn Jie, Wei Liu, and Jiebo Luo. 2018. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426* (2018).

[3] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 764–773.

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6569–6578.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. 2019. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *ACL*.

[9] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1115–1124.

[10] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4555–4564.

[11] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 787–798.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.

[13] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. 2020. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10880–10889.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[16] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-aware visual policy network for sequence-level image captioning. In *Proceedings of the 26th ACM international conference on Multimedia*. 1416–1424.

[17] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. 2019. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4673–4682.

[18] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. 2020. Learning cross-modal context graph for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11645–11652.

[19] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 11–20.

[20] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*. Springer, 792–807.

[21] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[23] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 2021. Structured Multi-Level Interaction Network for Video Moment Localization via Language Query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7026–7035.

[24] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. 2019. Neighbourhood watch: Referring expression comprehension via

language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1960–1968.

[25] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4644–4653.

[26] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. Improving one-stage visual grounding by recursive sub-query construction. *arXiv preprint arXiv:2008.01059* (2020).

[27] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4683–4693.

[28] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. 2018. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2403–2412.

[29] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1307–1315.

[30] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*. Springer, 69–85.

[31] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7282–7290.

[32] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2019. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[33] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13278–13288.

[34] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as points. *arXiv preprint arXiv:1904.07850* (2019).

[35] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4777–4786.

[36] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*. Springer, 391–405.