Propagating Over Phrase Relations for One-Stage Visual Grounding

Sibei Yang[1], Guanbin Li[2], and Yizhou Yu[1(✉)]

[1] The University of Hong Kong, Pokfulam, Hong Kong
sbyang9@hku.hk, yizhouy@acm.org
[2] Sun Yat-sen University, Guangzhou, China
liguanbin@mail.sysu.edu.cn

Abstract. Phrase level visual grounding aims to locate in an image the corresponding visual regions referred to by multiple noun phrases in a given sentence. Its challenge comes not only from large variations in visual contents and unrestricted phrase descriptions but also from unambiguous referrals derived from phrase relational reasoning. In this paper, we propose a linguistic structure guided propagation network for one-stage phrase grounding. It explicitly explores the linguistic structure of the sentence and performs relational propagation among noun phrases under the guidance of the linguistic relations between them. Specifically, we first construct a linguistic graph parsed from the sentence and then capture multimodal feature maps for all the phrasal nodes independently. The node features are then propagated over the edges with a tailor-designed relational propagation module and ultimately integrated for final prediction. Experiments on Flickr30K Entities dataset show that our model outperforms state-of-the-art methods and demonstrate the effectiveness of propagating among phrases with linguistic relations (Source code will be available at https://github.com/sibeiyang/lspn.).

Keywords: One-stage phrase grounding · Linguistic graph · Relational propagation · Visual grounding

1 Introduction

A fundamental yet challenging problem of AI for achieving communication between humans and machines in the real world is to perform jointly understanding of natural language and visual scene. To bridge language and vision, it is necessary to align visual contents in a given visual scene with the corresponding linguistic elements in the natural language which describes the visual scene. Phrase grounding [14], a basic task on language grounding to vision, has attracted increasing attention [3,12,19,29].

The phrase grounding is typically defined as locating corresponding visual regions in an image referred to by multiple noun phrases in a natural language

The first author is supported by the Hong Kong PhD Fellowship.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12364, pp. 589–605, 2020.
https://doi.org/10.1007/978-3-030-58529-7_35

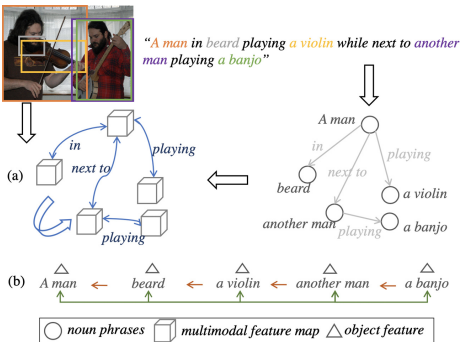

Fig. 1. Comparison of the proposed LSPN (a) with existing methods (b) on relational propagation among noun phrases for phrase grounding. LSPN (a) encodes language-vision information at nodes as multimodal feature maps. Then it propagates multimodal information over the parsed linguistic graph which encodes the linguistic structure. As a comparison, the existing methods (b) consider the object-level features and propagate the object information without considering explicit linguistic relations among phrases. They pass messages over all the pairs of phrases or sequential phrases following the reverse lexical order of the sentence.

description. Beyond object detection [15,17], a traditional vision task, phrase grounding introduces the natural language description and presents two extra challenges. First, phrase grounding generalizes the restricted object categories into unrestricted noun phrases description, which increases the difficulty for matching a separate noun phrase with a visual region due to the large variations in the pairs of object appearance and its related phrase description. Second, a noun phrase may only be able to unambiguously locate its corresponding visual region by cooperating with other specific phrases in the sentence. The noun phrases existing in a natural language description have phrase contexts, *i.e.*, the relations among phrases. For the sentence given in Fig. 1, its phrase contexts include relational triplets of "*A man-in-beard*", "*A man-playing-a violin*", "*A man-next to-another man*" and "*another man-playing-beard*". Note that there are two men and a unique violin in the image, the grounding result of the unique "*violin*" can be leveraged to distinguish the target man from the other man for the noun phrase "*A man*" by considering the relation of "*A man-playing-a violin*". Similarly, other phrase relations (*i.e.*, "*A man-in-beard*", "*A man-next to-another man*" and "*another man-playing-a banjo*") in the sentence can also help to ground noun phrases and refine the grounding results. Significantly, the indirect relations among phrases, *i.e.*, multi-order relations, may also be useful. For example, the relations of "*A man-in-beard*" and "*A man-playing to-a violin*" jointly help to identify the target beard for the noun phrase "*beard*".

However, most of existing works on phrase grounding ground noun phrases of a language description in an image individually without modeling the relations among phrases. They focus on learning the feature fusion in language and vision modalities [4,19,29], reconstructing the phrases from phrase-region features [7,

18] or matching the phrase embedding with the encoded phrase-related/phrase-unrelated region features [12,23] to address the first challenge mentioned above. There are few works taking phrase contexts into consideration, but they capture the partial or coarse phrase contexts without explicit linguistic relations among phrases (shown in Fig. 1(b)), including coreference relations [24], phrase-pair cues [1,13], contextual rewards [2] and sequential phrases following the reverse lexical order of the sentence [3].

To address the limitations mentioned above, we propose a Linguistic Structure guided Propagation Network (LSPN) for phrase grounding. The core ideas behind the proposed LSPN come from three aspects which include linguistic graph parsing from the input description, relational propagation for each pair of phrases with their relation, and one-stage grounding framework cooperated with iteratively relational propagation over the parsed linguistic graph. First, we parse the natural language description into a linguistic graph [27] and refine the graph based on the given noun phrases, where the nodes and edges of the graph are corresponding to the noun phrases and their relations respectively. The linguistic graph involves globally structured linguistic information, which also provides the possibility for indirectly relational propagation. Second, we propose a relational propagation module to perform message passing between a pair of subject and object phrases with their relation (*i.e.*, the relational triplet of *subject-relation-object*). Note that the relation between two phrases should be bidirectional, and the message from one phrase helps to unambiguously identify the corresponding visual region or refine the grounding result for the other phrase. Last but not least, we iteratively propagate language-vision multimodal information over the parsed linguistic graph to locate corresponding visual regions for the noun phrases in a single stage.

In summary, this paper has the following contributions:

- A relational propagation module (RPM) is proposed to perform bidirectional message passing for each pair of phrases with their linguistic relation.
- A linguistic structure guided propagation network is proposed for one-stage phrase grounding, which iteratively propagates the language-vision multimodal information for noun phrases using RPM under the guidance of parsed linguistic graph of the description.
- The experimental results on the common benchmark Flickr30K Entities dataset demonstrate that the proposed model outperforms state-of-the-art methods and shows the effectiveness of propagating over phrase relations.

2 Related Work

Phrase Grounding. Building a direct connection between textual phrases and visual contents is necessary for phrase grounding. Some works first fuse the representations in vision and language modalities, and then predict the visual regions [19,29] or learn the multimodal similarities for pairs of phrases and visual regions [4,22]. Another works [7,18] address phrase grounding from the view of

phrase construction. Plummer et al. [12] group phrases into different sets and learn the group-conditional embeddings for phrases.

However, the above works treat phrases in isolation and neglect relations among them. Wang et al. [24] focus on one specific type of relations between phrases, *i.e.*, coreference relations (*e.g.*, "man" and "his hand"), and learn the structured matching with relation constraints. Plummer et al. [13] perform joint inference over phrases during test stage by combining extracted image and language cues, and they only consider the phrase-pair spatial cues. The works [2] and [3] implicitly consider phrase contexts, the former refines grounding results by using contextual information from all other phrases as rewards, and the latter sequentially predicts the grounding results for the phrases following their reverse lexical order in the sentence.

Different from existing methods, we explicitly extract the relations between phrases by parsing the linguistic structure of the sentence and propagate over phrase relations to build the interactions among phrases.

Referring expression comprehension aims to locate in an image a visual object described by a natural language expression. Recent works on it also try to explore the relational contexts for objects to help distinguish the referent from other objects. Yang et al. [25,28] encode the expression-guided multi-order relations by performing a gated graph convolutional networks over a multimodal relation graph based on objects in the image. Some works [6,26,31] capture the context-related language information by using self-attention mechanism over words in the expressions. In particular, Yu et al. [31] compute the matching scores between the attended relation embedding and the referent's relative location differences with its surrounding objects to capture the relational context. Yang et al. [26] highlight language information about objects and relations in a stepwise manner, and locate its corresponding visual evidence in the image.

However, most of the existing works on referring expression comprehension also neglect the syntax of the referring expression and only consider very limited contextual relations. Yang et al. [27] and Liu et al. [10] use the parsed linguistic structure of the expression to guide the process of locating the referent, but the relation models they build are not very suitable for phrase grounding. Specifically, noun phrases, except referent phrase, and their relations are used to modify the referent, and the process of locating the referent is from bottom to up. Instead of finding the referent, the aim of phrase grounding is to ground all the noun phrases in the sentence, and every noun phrase deserves attention. Thus, the relations between noun phrases on phrase grounding should be bidirectional.

Single-stage networks for object detection are widely used due to their fast inference speed and high accuracy. Recently, single-stage grounding networks have been proposed for phrase grounding. Yeh et al. [30] minimizes the energy based on a set of visual concepts over a large number of bounding boxes. However, the visual concepts used by it are based on multiple extra pre-trained models, and it is not clear how to optimize the visual concepts and the grounding

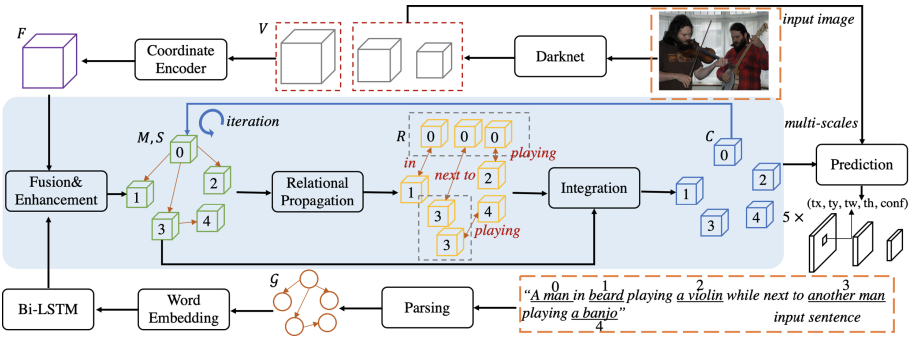

Fig. 2. The overall architecture of the proposed LSPN for one-stage phrase grounding. First, LSPN encodes the input image as spatial-aware feature map F by fusing the visual feature map V with spatial coordinates P. Second, the linguistic graph $\mathcal{G}$ is parsed from the input sentence. Third, for each node, LSPN obtains the multimodal feature map M and phrase-conditional enhance map S from the language representation of node and spatial-aware feature map. Next, LSPN captures relational enhance map R and combined enhance map C by passing messages over edges using relational propagation module and integrating messages for nodes. The propagation can be performed multiple times. Finally, LSPN predicts the grounding results from the final combined feature map.

model from end to end. Yang et al. [29] and Sadhu et al. [19] directly fuse the language feature of the input phrase and the spatial features of the image feature maps into single-stage object detection frameworks, *i.e.*, YOLOv3 [16] and SSD [11], respectively. However, existing one-stage grounding approaches ignore the fact that the referential meaning of noun phrases may depend on other word components of the sentence. Thus, we propose a one-stage grounding network which allows relational propagation between phrases under the guidance of the linguistic structure of the sentence.

3 Approach

The proposed linguistic structure guided propagation network (LSPN) iteratively propagates the language-vision multimodal information among noun phrases under the guidance of a linguistic graph parsed from a natural language description and grounds noun phrases corresponding visual regions in an image. The framework of LSPN is illustrated in Fig. 2, and it consists of three main modules, *i.e.*, image and language representation, relational propagation and prediction.

3.1 Image and Language Representation

We represent an input image and a natural language description as spatial-aware feature maps and a linguistic graph respectively. The spatial-aware feature maps

capture the global image contexts. They are obtained by fusing the visual feature maps extracted from a CNN backbone with spatial coordinates embedding. The linguistic graph encodes the description's linguistic structure and provides the guidance for relational propagation among noun phrases.

Image Encoder. The proposed one-stage LSPN is based on the YOLOv3 [16] object detection framework, and we adopt the Darknet-53 [16] with feature pyramid networks [8] as the visual feature extractor. Following [29], we resize the input image I to 256×256 with zero padding and keep its aspect ratio, and extract the outputs of feature pyramid networks as visual feature maps with spatial resolutions and channels of $8 \times 8 \times 1024$, $16 \times 16 \times 512$ and $32 \times 32 \times 256$, respectively. To simplify writing, we denote a feature map with the size of $W \times H \times D_v$ as V to introduce the computations of LSPN.

A noun phrase may describe not only the appearance of a visual region itself but also its location in the image, such as "*right man*" and "*the bottle in the middle*". Thus, similar to previous methods [19,29], we embed the spatial coordinates of a feature map into the visual features to form a spatial-aware version. In particular, the spatial map P is of the same spatial resolution as its corresponding visual feature map V, *i.e.*, $W \times H$, and the spatial feature at each position $(x, y) \in \{(0,0), (0,1), ..., (W-1, H-1)\}$ is defined as,

$$P_{x,y} = [\frac{x}{W}, \frac{y}{H}, \frac{x+0.5}{W}, \frac{y+0.5}{H}, \frac{x+1}{W}, \frac{y+1}{H}, \frac{1}{W}, \frac{1}{H}], \tag{1}$$

where the vector $P_{x,y} \in \mathbb{R}^8$ encodes the normalized coordinates of top-left, center, bottom-right, width and height of the grid at position (x, y). Next, we fuse the visual feature map V with the spatial map P to obtain the spatial-aware feature map $F \in \mathbb{R}^{W \times H \times D_f}$,

$$F = [\text{L2Norm}(\text{Conv}_0(V)); P]. \tag{2}$$

where the $\text{Conv}_0(\cdot)$ is a convolutional layer with kernel size 1×1, $\text{L2Norm}(\cdot)$ is the L2 normalization over the feature channel, and $[;]$ refers to the concatenation operation.

Linguistic Graph Parsing. The linguistic graph encodes the description as a graph where the nodes and edges respectively correspond to the noun phrases and the linguistic relations (*i.e.*, preposition or verb phrases) between noun phrases mentioned in the description. We construct the linguistic graph by parsing the description as an initial scene graph and then refining the initial scene graph based on given noun phrases. Given a natural language description L and a set of noun phrases $\mathcal{P}_g$ in L, the construction process for the linguistic graph $\mathcal{G}$ is summarized as follows,

– We first parse the natural language description L into an initial scene graph [27] using an off-the-shelf scene graph parser [20]. The nodes and edges of the initial scene graph correspond to nouns with modifiers (*e.g.*, determinants and adjectives) and linguistic relations between nouns in L.

- Then, for each node, we reorganize it as a noun phrase by sorting the noun and its modifiers following their original order in the description. The set of reorganized noun phrases is denoted as $\mathcal{P}_r$.
- However, the given noun phrases $\mathcal{P}_g$ in the description L sometimes may not exactly match with the noun phrases $\mathcal{P}_r$ in the parsed scene graph. Therefore, we associate each given noun phrase with one parsed noun phrase which has maximum overlap words with the given noun phrase. Then, we replace the parsed noun phrase by the given noun phrase.
- Next, for each parsed edge, we further insert or delete the words in it based on the replaced noun phrases connected by it. Finally, we obtain the resulted linguistic graph $\mathcal{G}$ from the refined edges and noun phrases in the scene graph.

The linguistic graph $\mathcal{G}$ parsed from the language description L is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_n\}_{n=1}^{N}$ is a set of nodes and $\mathcal{E} = \{e_k\}_{k=1}^{K}$ is a set of directed edges. Specifically, each node v_n corresponds to a noun phrase L_n with a sequence of words in L, and each edge e_k is a triplet $e_k = (e_k^{(s)}, e_k^{(r)}, e_k^{(o)})$. In the triplet, $e_k^{(s)} \in \mathcal{V}$ and $e_k^{(o)} \in \mathcal{V}$ are the subject node and the object node respectively, and $e_k^{(r)}$ associating with a preposition or verb phrase E_k in L is the linguistic relation from $e_k^{(s)}$ to $e_k^{(o)}$. In addition, we adopt $\mathcal{E}_n^{\text{in}} \subset \mathcal{E}$ to denote the set of edges whose object node is v_n, use $\mathcal{E}_n^{\text{out}} \subset \mathcal{E}$ to denote the set of edges whose subject node is v_n and denote de_n as the degree of node v_n.

3.2 Relational Propagation

The proposed relational propagation is implemented by passing messages at individual nodes over the parsed linguistic graph $\mathcal{G}$. We first obtain the relation-unrelated multimodal features for all the nodes $\mathcal{V}$ independently, and then propagate them over all the edges $\mathcal{E}$ by considering each edge separately and integrate the passed information for nodes. In particular, the bidirectional propagation over a single edge is achieved by the relational propagation module.

Propagation over Linguistic Graph. We first obtain the multimodal features for all the nodes $\mathcal{V}$ in graph $\mathcal{G}$ by fusing the spatial-aware feature map F mentioned in Sect. 3.1 with the language representations of noun phrases at nodes. In particular, we encode each word as a word embedding vector, and the initial phrasal embedding at a node is set to the mean pooling of the embedding vectors of all the words in the phrase. For a node v_n with noun phrase L_n and its phrasal embedding vector $w_n \in \mathbb{R}^{D_w}$, we learn its phrase feature $w_n^{'} \in \mathbb{R}^{D_{w'}}$ from initial phrasal embedding via a nonlinear transformation,

$$w_n^{'} = \text{L2Norm}(\text{MLP}_0(w_n)), \tag{3}$$

where the $\text{MLP}_0(\cdot)$ consists of multiple linear layers with ReLU activation functions, and the $\text{L2Norm}(\cdot)$ is the L2 normalization. Next, we obtain a multimodal feature map $M_n \in \mathbb{R}^{W \times H \times D_m}$ by fusing the phrase feature $w_n^{'}$ with the feature

map F and meanwhile learn a phrase-conditional enhance map $S_n \in \mathbb{R}^{W\times H\times D_s}$, which is formulated as,

$$\begin{aligned} M_n &= \text{L2Norm}(\text{Conv}_1([F; \text{Tile}(w_n^{'})])), \\ S_n &= \sigma(\text{Conv}_3(\text{Conv}_2(F) + \text{Tile}(\text{Fc}_0(w_n^{'})))), \end{aligned} \tag{4}$$

where $\text{Tile}(\cdot)$ is to tile a vector to each spatial position of a feature map with resolution $W \times H$, $\text{Conv}_1(\cdot)$ is a series of convolutional layers along with BatchNorm and ReLU, $\text{Fc}_0(\cdot)$ is a fully connected layer, $\sigma(\cdot)$ is the sigmoid activation, and $\text{Conv}_2(\cdot)$ and $\text{Conv}_3(\cdot)$ are two convolutional layers with kernel size 1×1.

After obtaining the multimodal information for all the nodes $\mathcal{V}$, we pass it over the edges $\mathcal{E}$ in linguistic graph $\mathcal{G}$. For an edge $e_k = (e_k^{(s)}, e_k^{(r)}, e_k^{(o)})$, we first encode its linguistic feature. Specifically, we integrate the phrases associated with $e_k^{(s)}$, $e_k^{(r)}$ and $e_k^{(o)}$ as a sequence and pass the word embedding vectors in the sequence into a bidirectional LSTM [5], and the linguistic feature is the concatenation of the last hidden states of both the forward and backward LSTMs. The linguistic feature is denoted as h_k. Then, we feed its linguistic feature and multimodal information (*i.e.*, the multimodal feature maps and the phrase-conditional enhance maps) at subject node $e_k^{(s)}$ and object node $e_k^{(o)}$ into the **relational propagation module** to obtain the relational enhance maps, which are denoted as $R_k^{(s)} \in \mathbb{R}^{W\times H\times D_s}$ and $R_k^{(o)} \in \mathbb{R}^{W\times H\times D_s}$.

Next, for each node v_n, we integrate the relational enhance maps obtained from edges in the sets $\mathcal{E}_n^{\text{out}}$ and $\mathcal{E}_n^{\text{in}}$ to get the final relational enhance map, and further combine it with the initial phrase-conditional enhance map S_n. The combined enhance map $C_n \in \mathbb{R}^{W\times H\times D_s}$ at node v_n is computed as follows,

$$\begin{aligned} R_n &= \frac{\sum_{e_{k'}\in\mathcal{E}_n^{\text{out}}} R_{k'}^{(s)} + \sum_{e_{k''}\in\mathcal{E}_n^{\text{in}}} R_{k''}^{(o)}}{de_n} \\ C_n &= \begin{cases} S_n, & \text{if } de_n = 0, \\ (S_n + R_n)/2, & \text{otherwise}, \end{cases} \end{aligned} \tag{5}$$

where de_n is the degree of node v_n (defined in Sect. 3.1).

Note that we can iteratively perform the propagation over the linguistic graph multiple times. At each time step, we can use the combined enhance maps at the last time step to replace the phrase-conditional enhance maps as the inputs of the relational propagation module to update the combined enhance maps. Iterative propagation can help to capture indirect relations among nodes. For each node v_n, M_n is the fundamental multimodal feature map and is not changed during each iterative relational propagation. S_n, which is used to enhance M_n, is updated after each iterative relational propagation. At each time step, M_n is replaced as the combined enhance map C_n of the last time step.

Relational Propagation Module. The relational propagation module passes the message of a single edge over its pair of nodes under the guidance of its

linguistic feature and outputs the relational enhance maps for the nodes. Note that although the edge from the subject node to the object node is directed, the relational propagation between the subject node and object node should be bidirectional as the message from one node helps to unambiguously ground and refine the result for the other phrase.

Given the multimodal feature map M_{sub} and the phrase-conditional enhance map S_{sub} at subject node v_{sub}, M_{obj} and S_{obj} at object node v_{obj} and the edge's linguistic feature h, the relational enhance map $R_{sub} \in \mathbb{R}^{W \times H \times D_s}$ for the subject node is computed as follows,

$$\begin{aligned} g_{obj} &= \text{MLP}_{obj}([\text{AvgPool}(M_{obj} \circ S_{obj}); h]), \\ M^{'}_{sub} &= \text{Conv}_{sub0}(M_{sub} \circ S_{sub}), \\ R_{sub} &= \sigma(\text{Conv}_{sub1}(\gamma(M^{'}_{sub} + \text{Tile}(g_{obj})))), \end{aligned} \tag{6}$$

where $\text{MLP}_{obj}(\cdot)$ is a multi-layer perceptron, $\text{AvgPool}(\cdot)$ means the global average pooling, $\circ$ represents element-wise multiplication, $\text{Conv}_{sub0}(\cdot)$ and $\text{Conv}_{sub1}(\cdot)$ are two convolutional layers and γ refers to the ReLU activation function. S_{obj} and S_{sub} are used to enhance M_{obj} and M_{sub}, respectively. g_{obj} provides the relational guidance for subject node, and it encodes the linguistic feature of edge and the global multimodal feature from object node.

Moreover, the relational enhance map $R_{obj} \in \mathbb{R}^{W \times H \times D_s}$ for the object node can be obtained following the similar computation.

3.3 Prediction and Loss

The prediction of phrase grounding is similar to the bounding boxes detection in YOLOv3 [16]. Following [29], we match three anchor boxes to every spatial position of a feature map, choose the candidate box with highest confidence score over all the anchor boxes of three feature maps at various spatial resolutions, and obtain the final grounding result by regressing the candidate box using the predicted regression offsets.

For each node v_n in graph $\mathcal{G}$, the regression offsets and confidence scores $pred_n \in \mathbb{R}^{W \times H \times 15}$ for the three anchor boxes at a single spatial resolution $W \times H$ are computed as follows,

$$pred_n = \text{Conv}_{pred}(M_n \circ C_n), \tag{7}$$

where the multimodal feature map M_n and the final combined enhance map C_n are mentioned in Sect. 3.1 and $\text{Conv}_{pred}(\cdot)$ is a series of convolutional layers.

During training, we compute two types of losses (*i.e.*, a classification cross-entropy loss $Loss_{conf}$ and a L1 regression loss $Loss_{reg}$) and combine them as the final loss,

$$Loss = Loss_{conf} + \lambda Loss_{reg}, \tag{8}$$

where λ is used to balance the $Loss_{conf}$ and $Loss_{reg}$. In particular, the classification loss $Loss_{conf}$ is the cross entropy loss between the output of a softmax

function over all anchor boxes of three feature maps at various spatial resolutions and an one-hot vector labeling the anchor box with highest Intersection over Union (IoU) with the ground truth region set as 1. And the regression loss $Loss_{reg}$ is the L1 loss between the predicted regression offsets and the target regression offsets. Specifically, the target regression offsets $\mathbf{t} = [t_x, t_y, t_w, t_h] \in \mathbb{R}^4$ are defined as,

$$t_x = (g_x - r_x)/r_w, \qquad t_y = (g_y - r_y)/r_h, \tag{9}$$

$$t_w = \log(g_w/r_w), \qquad t_h = \log(g_h/r_h), \tag{10}$$

where $\mathbf{g} = [g_x, g_y, g_w, g_h] \in \mathbb{R}^4$ and $\mathbf{r} = [r_x, r_y, r_w, r_h] \in \mathbb{R}^4$ are the coordinates of the ground truth box and the candidate box, respectively.

During inference, we obtain the predicted box $\hat{\mathbf{g}} = [\hat{g}_x, \hat{g}_y, \hat{g}_w, \hat{g}_h]$ based on the chosen box $\mathbf{r}$ and the predicted regression offsets $\mathbf{t}'$,

$$\hat{g}_x = r_w * t'_x + r_x, \qquad \hat{g}_y = r_h t'_y + r_y, \tag{11}$$

$$\hat{g}_w = r_w \exp(t'_w), \qquad \hat{g}_h = r_h \exp(t'_h). \tag{12}$$

4 Experiments

4.1 Dataset and Evaluation

Dataset. We have conducted experiments on the commonly used Flickr30K Entities dataset [14] for phrase grounding. The phrase contexts in a natural language description are considered for bounding box annotations on Flickr30K Entities dataset. In Flickrr30K, a single noun phrase may be associated with multiple ground truth bounding boxes, while a single bounding box can also be matched with multiple noun phrases. Following previous works [3,29], if a noun phrase has multiple ground truth bounding boxes, it will be associated with the union of its all correlated boxes. We adopt the same training, validation and test split used in previous methods [3,29].

Evaluation Metric. The grounding *accuracy* is adopted as the evaluation metric, which is defined as the fraction of correct predictions for noun phrases grounding, and one prediction is considered correct if the IoU between the predicted bounding box and the ground truth region is larger than 0.5. Besides, the *inference speed* is important for models in real-time applications. The inference time is also reported, and all the tests are conducted on a desktop with the Intel Xeon Gold 5118@2.30GHz and NVIDIA RTX 2080TI.

4.2 Implementation

We extract visual feature maps from the Darknet-53 [16] with feature pyramid networks [8] pre-trained on MSCOCO object detection [9] following previous one-stage model [29]. The channel dimension of a spatial-aware feature map is set to

1024 (*i.e.*, $D_f = 1024$). The dimension of the hidden state of the bidirectional LSTM is set to 512. Thus, the linguistic features of edges are 1024-dimensional vectors, *i.e.*, $D_h = 1024$. The remaining hyper-parameters about the feature dimensions are set to 512. The RMSProp optimizer [21] is adopted to update network parameters, and the learning rate is initially set to 1e-4 and decreases following a polynomial schedule with power of 1. The learning rate for learnable parameters in Darknet-53 is set to one-tenth of the main learning rate. The loss balancing factor λ is set to 5. We train the model for 140k iterations with the batch size set to 16.

4.3 Comparison with the State of the Art

We evaluate the proposed LSPN on the Flickr30K Entities dataset and compare it with state-of-the-art methods. The results are shown in Table 1, LSPN achieves the best performance at 69.53% in accuracy and outperforms all the state-of-the-art models. It improves the accuracy achieved by the existing best performing method by 1.91%, which demonstrates the effectiveness of propagation over phrase relations in LSPN.

Table 1. Comparison with the state-of-the-art methods on Flickr30K Entities w.r.t accuracy metric and inference time for one image-query pair. We use * to indicate one-stage models. None-superscript indicates that model is from a two-stage method. The best performing method is marked in bold.

Method	Accuracy (%)	Time (ms)
GroundeR [18]	47.81	-
RtP [14]	50.89	-
IGOP [30]	53.97	-
SPC+PPC [13]	55.49	-
SS+QRN [2]	55.99	-
SimNet-ResNet [22]	60.89	140
CITE-ResNet [12]	61.33	149
SeqGROUND [3]	61.60	-
ZSGNet* [19]	63.39	-
G^3RAPH++ [1]	66.93	-
FAOS* [29]	67.62	**16**
Ours LSPN*	**69.53**	20

As shown in the rightmost column of Table 1, the inference speed of one-stage methods (*i.e.*, FAOS and ours LSPN) is much faster than that of the two-stage methods (*i.e.*, SimNet-ResNet and CITE-ResNet). It takes the two-stage methods generally more than 140ms to ground a language query in an

image. Most of the time is spent on generating region proposals in the image and extracting features for them. In contrast, the one-stage methods take less than 20ms to process one image-query pair without generating region proposals. Compared to FAOS, the proposed LSPN propagating contexts over noun phrases with linguistic relations achieves a higher grounding accuracy, though at the expense of a little bit of time cost.

Table 2. Comparison over coarse categories on Flickr30K Entities using accuracy metric (in percentage). The best performing method is marked in bold.

Method	People	Clothing	Body parts	Animals	Vehicles	Instruments	Scene	Other
SMPL	57.89	34.61	15.87	55.98	52.25	23.46	34.22	26.23
GroundeR	61.00	38.12	10.33	62.55	68.75	36.42	58.18	29.08
RtP	64.73	46.88	17.21	65.83	68.72	37.65	51.39	31.77
IGOP	68.71	56.83	19.50	70.07	73.72	39.50	60.38	32.45
SPC+PPC	71.69	50.95	25.24	76.23	66.50	35.80	51.51	35.98
CITE	73.20	52.34	30.59	76.25	75.75	48.15	55.64	42.83
SeqGROUND	76.02	56.94	26.18	75.56	66.00	39.36	68.69	40.60
G^3RAPH++	78.86	**68.34**	39.80	**81.38**	76.58	42.35	68.82	45.08
Ours LSPN	**80.69**	67.17	**44.17**	79.92	**83.23**	**62.96**	**70.91**	**52.82**

Moreover, we provide the phrase grounding performance over coarse categories. As shown in Table 2, LSPN consistently surpasses all the state-of-the-art methods on six categories, and achieves consistent improvement in overall accuracy over all the categories compared to all other methods. It significantly improves the accuracy on categories of instruments, vehicles, body parts and other by 14.81%, 7.07%, 4.37% and 7.74% respectively.

4.4 Ablation Study

We conduct an ablation study on the proposed LSPN to demonstrate the effectiveness and necessity of each component and have trained six additional variants of our model for comparison. The results are shown in Table 3.

- The *multimodal* model is the baseline, which predicts the confidence scores and regression offsets of each anchor box from the multimodal feature maps that are incorporated with the visual feature, spatial information and phrase feature.
- The *enhance* model extends the multimodal model by using the phrase-conditional enhance maps to enhance the multimodal feature maps, which improves the performance by 0.31% in accuracy.
- The *linguistic graph propagation(1)* model performs the relational propagation over the linguistic graph once. It achieves the best accuracy of 69.53% among the seven models and improves the accuracy by 1.38% over that achieved by the multimodal model, which demonstrates the effectiveness of

Table 3. Ablation study on variances of the proposed LSPN on Flickr30K Entities using accuracy metric. The number in parentheses refers to the number of propagation steps in our model.

Method	Accuracy (%)
Multimodal	68.15
Enhance	68.46
Linguistic graph propagation(1)	69.53
Linguistic graph propagation(2)	69.52
Subject graph propagation(1)	68.81
Object graph propagation(1)	68.97
Contextual propagation(1)	67.14

considering the relational propagation between noun phrases. The *linguistic graph propagation(2)* model is similar to linguistic graph propagation(1) model but propagates over the phrase relations twice. It does not further improve the performance and achieves similar accuracy as lingtuistic graph propagation(1). The reason may be that the number of phrases that need to rely on indirect phrase relations to be unambiguously grounded accounts for a relatively small proportion, and multiple propagations may instead introduce context noise.

- The *subject graph propagation(1)* model and the *object graph propagation(1)* model perform one-way subject-to-object and object-to-subject propagation over noun phrases. Compared to linguistic graph propagation(1) model performing bidirectional propagation, the performance of subject graph propagation(1) model and the object graph propagation(1) model is worse than that of it in accuracy by 0.72% and 0.56% respectively. The results demonstrate that the importance of bidirectional message passing for pairs of noun phrases.
- The *contextual propagation* model explores the message passing over another constructed graph without the explicit guidance of the linguistic structure. For each noun phrase in a sentence, we separately connect its three nearest noun phrases as three edges and learns relational weights for these edges by using the global context of the sentence. We then perform relational propagation over the constructed graph which is built on noun phrase and edges with learned weights and evaluate on this algorithmic variant. The worse experimental performance has demonstrated that the propagation over incompletely correct relations may adversely affect the model, and adopting the parsed linguistic graph as guidance is crucial for relational propagation.

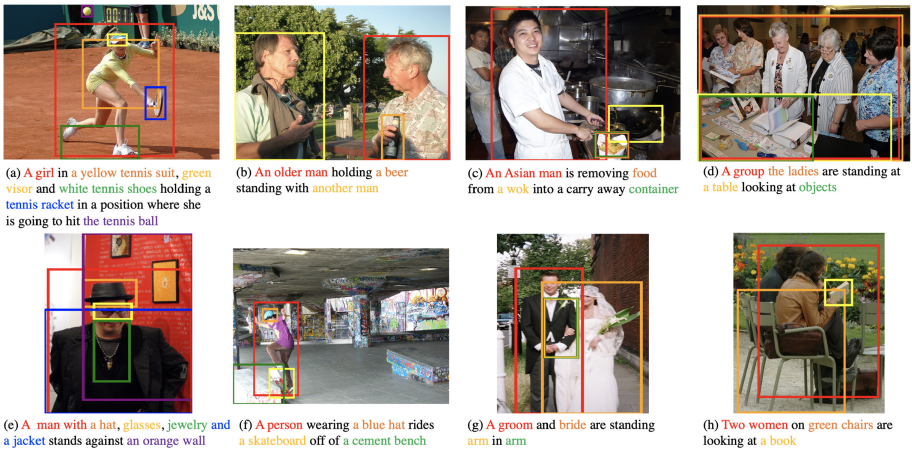

Fig. 3. Qualitative results showing noun phrases in sentences and their grounding results predicted by LSPN.

4.5 Qualitative Evaluation

The qualitative evaluation results for phrase grounding are shown in Fig. 3. The proposed LSPN is able to successfully locate the visual regions referred to by noun phrases in different kinds of challenging scenarios.

In (a) and (e), LSPN grounds multiple noun phrases in long sentences, and it correctly identifies the corresponding object for each phrase. In (b), (c), (f) and (h), LSPN unambiguously distinguishes the referred objects from other objects belonging to the same categories by considering their relations to other objects in the sentence. For the example in (b), "*one older man*" can be identified by considering its relation ("*holding*") to "a beer". Samples (c) and (g) show that a single object in the image can be referred by multiple noun phrases. In (d) and (h), a noun phrase may be associated with multiple visual objects, LSPN is able to successfully locate them from the single noun phrase. For the example in (h), LSPN finds the two "*green chairs*" while excludes the chair on the right.

5 Conclusions

In this paper, we have proposed a linguistic structure guided propagation network (LSPN) for one-stage phrase grounding. LSPN works by iteratively propagating the language-vision multimodal information between noun phrases under the guidance of the linguistic graph and locating the image region corresponding to each noun phrase in the referring sentence. The context relation between each pair of noun phrases is captured by a relational propagation module. Experimental results on the common benchmark Flickr30K Entities dataset demonstrate that the proposed model outperforms state-of-the-art methods and shows the effectiveness of propagating over phrase relations.

Acknowledgment. This work is partially supported by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2020B1515020048 and the National Natural Science Foundation of China under Grant No. U1811463.

References

1. Bajaj, M., Wang, L., Sigal, L.: G3raphground: graph-based language grounding. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 2019
2. Chen, K., Kovvuri, R., Nevatia, R.: Query-guided regression network with context policy for phrase grounding. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 824–832 (2017)
3. Dogan, P., Sigal, L., Gross, M.: Neural sequential phrase grounding (seqground). In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4175–4184 (2019)
4. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2016)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
6. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1115–1124 (2017)
7. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4555–4564 (2016)
8. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
9. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
10. Liu, D., Zhang, H., Zha, Z.J., Feng, W.: Learning to assemble neural module tree networks for visual grounding. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
11. Liu, W., et al.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
12. Plummer, B.A., Kordas, P., Kiapour, M.H., Zheng, S., Piramuthu, R., Lazebnik, S.: Conditional image-text embedding networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11216, pp. 258–274. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_16
13. Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik, S.: Phrase localization and visual relationship detection with comprehensive image-language cues. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1928–1937 (2017)

14. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer VIsion, pp. 2641–2649 (2015)
15. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
16. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
18. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 817–834. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_49
19. Sadhu, A., Chen, K., Nevatia, R.: Zero-shot grounding of objects from natural language queries. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4694–4703 (2019)
20. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: Workshop on Vision and Language (VL15). Association for Computational Linguistics, Lisbon, Portugal, September 2015
21. Tieleman, T., Hinton, G.: Lecture 6.5–RmsProp: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning (2012)
22. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 394–407 (2018)
23. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013 (2016)
24. Wang, M., Azab, M., Kojima, N., Mihalcea, R., Deng, J.: Structured matching for phrase localization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 696–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_42
25. Yang, S., Li, G., Yu, Y.: Cross-modal relationship inference for grounding referring expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4145–4154 (2019)
26. Yang, S., Li, G., Yu, Y.: Dynamic graph attention for referring expression comprehension. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4644–4653 (2019)
27. Yang, S., Li, G., Yu, Y.: Graph-structured referring expression reasoning in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
28. Yang, S., Li, G., Yu, Y.: Relationship-embedded representation learning for grounding referring expressions. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
29. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4683–4693 (2019)

30. Yeh, R., Xiong, J., Hwu, W.M., Do, M., Schwing, A.: Interpretable and globally optimal prediction for textual grounding using image concepts. In: Advances in Neural Information Processing Systems, pp. 1912–1922 (2017)
31. Yu, L., et al.: MattNet: modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1307–1315 (2018)