Learning Cross-Modal Context Graph for Visual Grounding

Yongfei Liu1∗ Bo Wan1∗ Xiaodan Zhu2 Xuming He1
1 ShanghaiTech University 2 Queen’s University
{liuyf3, wanbo, hexm}@shanghaitech.edu.cn xiaodan.zhu@queensu.ca

Abstract
Visual grounding is a ubiquitous building block in many vision-language tasks and yet remains challenging due to large variations in visual and linguistic features of grounding entities, strong context effect and the resulting semantic ambiguities. Prior works typically focus on learning representations of individual phrases with limited context information. To address their limitations, this paper proposes a language-guided graph representation to capture the global context of grounding entities and their relations, and develop a cross-modal graph matching strategy for the multiple-phrase visual grounding task. In particular, we introduce a modular graph neural network to compute context-aware representations of phrases and object proposals respectively via message propagation, followed by a graph-based matching module to generate globally consistent localization of grounding phrases. We train the entire graph neural network jointly in a two-stage strategy and evaluate it on the Flickr30K Entities benchmark. Extensive experiments show that our method outperforms the prior state of the arts by a sizable margin, evidencing the efficacy of our grounding framework. Code is available at https://github.com/youngfly11/LCMCG-PyTorch.

1 Introduction
Integrating visual scene and natural language understanding is a fundamental problem toward achieving human-level artificial intelligence, and has attracted much attention due to rapid advances in computer vision and natural language processing (Mogadala, Kalimuthu, and Klakow 2019). A key step in bridging vision and language is to build a detailed correspondence between a visual scene and its related language descriptions. In particular, the task of grounding phrase descriptions into their corresponding image has become an ubiquitous building block in many vision-language applications, such as image retrieval (Justin et al. 2015; Nam et al. 2019), image captioning (Li et al. 2017; Feng et al. 2019), visual question answering (Mun et al. 2018; Cadene et al. 2019) and visual dialogue (Das et al. 2017; Kottur et al. 2018).

Many existing works on visual grounding tackle the problem by localizing each noun phrase independently via phrase-object matching (Plummer et al. 2015; Plummer et al. 2018; Yu et al. 2018b; Rohrbach et al. 2016). However, such grounding strategy tends to ignore visual and linguistic context, thus leading to matching ambiguity or errors for complex scenes. Only a few grounding approaches take into account context information (Pelin, Leonid, and Markus 2019; Chen, Kovvuri, and Nevatia 2017) or phrase relationship (Wang et al. 2016; Plummer et al. 2017) when representing visual or phrase entities. While they partially alleviate the problem of grounding ambiguity, their context or relation representations have several limitations for capturing global structures in language descriptions and visual scenes. First, for language context, they typically rely on chain-structured LSTMs defined on description sentences, which have difficulty in encoding long-range dependencies among phrases. In addition, most methods simply employ off-the-shelf object detectors to generate object candidates for cross-modal matching. However, it is inefficient to encode visual context for those objects due to a high ratio of false positives in such object proposal pools. Furthermore, when incorporating phrase relations, these methods often adopt a stage-wise strategy that learns representations of noun phrases and their relationship separately, which is sub-optimal for the overall grounding task.

In this work, we propose a novel cross-modal graph network to address the aforementioned limitations for multiple-
phrase visual grounding. Our main idea is to exploit the language description to build effective global context representations for all the grounding entities and their relations, which enables us to generate a selective set of high-quality object proposals from an image and to develop a context-aware cross-modal matching strategy. To achieve this, we design a modular graph neural network consisting of four main modules: a backbone network for extracting basic language and visual features, a phrase graph network for encoding phrases in the sentence description, a visual object graph network for computing object proposal features and a graph similarity network for global matching between phrases and object proposals.

Specifically, given an image and its textual description, we first use the backbone network to compute the language embedding for the description, and to generate an initial set of object proposals. To incorporate language context, we construct a language scene graph from the description (e.g., Schuster et al. 2015; Wang et al. 2018b) in which the nodes are noun phrases, and the edges encode relationships between phrases. Our second module, phrase graph network, is defined on this language scene graph and computes a context-aware phrase representation through message propagation on the phrase graph. We then use the phrase graph as a guidance to build a visual scene graph, in which the nodes are object proposals relevant to our phrases, and the edges encode the same type of relations as in the phrase graph between object proposals. The third network module, visual object graph network, is defined on this derived graph and generates a context-aware object representation via message propagation. Finally, we introduce a graph similarity network to predict the global matching of those two graph representations, taking into account similarities between both graph nodes and relation edges.

We adopt a two-stage strategy in our model learning, of which the first stage learns the phrase graph network and visual object features while the second stage trains the entire deep network jointly. We validate our approach by extensive experiments on the public benchmark Flickr30K Entities (Plummer et al. 2015), and our method outperforms the prior state of the art by a sizable margin. To better understand our method, we also provide the detailed ablative study of our context graph network.

The main contributions of our work are three-folds:

- We propose a language-guided graph representation, capable of encoding global contexts of phrases and visual objects, and a globally-optimized graph matching strategy for visual grounding.
- We develop a modular graph neural network to implement the graph-based visual grounding, and a two-stage learning strategy to train the entire model jointly.
- Our approach achieves new state-of-the-art performance on the Flickr30K Entities benchmark.

2 Related Works

Visual Grounding: In general, visual grounding aims to localize object regions in an image corresponding to multiple noun phrases from a sentence that describes the underlying scene. Rohrbach et al. (2016) proposed an attention mechanism to attend to relevant object proposals for a given phrase and designed a loss for phrase reconstruction. Plummer et al. (2018) presented an approach to jointly learn multiple text-conditioned embedding in a single end-to-end network. In DDPN (Yu et al. 2018b), they learned a diversified and discriminate proposal network to generate higher quality object candidates. Those methods grounded each phrase independently, ignoring the context information in image and language. Only a few approaches attempted to solve visual grounding by utilizing context cues. Chen et al. (2017) designed an additional reward by incorporating context phrases and train the whole network by reinforcement learning. Dongan et al. (2019) took context into account by adopting chain-structured LSTMs network to encode context cues in language and image respectively. In our work, we aim to build cross-modal graph networks under the guidance of language structure to learn global context representation for grounding entities and object candidates.

Referring Expression: Referring expression comprehension is closely related to visual grounding task, which attempts to localize expressions corresponding to image regions. Unlike visual grounding, those expressions are typically region-level descriptions without specifying grounding entities. Nagaraja et al. (2016) proposed to utilize LSTMs to encode visual and linguistic context information jointly for referring expression. Yu et al. (2018a) developed modular attention network, which utilized language-based attention and visual attention to localize the relevant regions. Wang et al. (2019) applied self-attention mechanism on sentences and built a directed graph over neighbour objects to model their relationships. All the above-mentioned methods fail to explore the structure of the expression explicitly. Our focus is to exploit the language structure to extract cross-modal context-aware representations.

Structured Prediction: Structured prediction is a framework to solve the problems whose output variables are mutually dependent or constrained. Justin et al. (2015) proposed the task of scene graph grounding to retrieve images, and formulated the problem as structured prediction by taking into account both object and relationship matching. To explore the semantic relations in visual grounding task, Wang et al. (2016) tried to introduce a relational constraint between phrases, but limited their relations to possessive pronouns only. Plummer et al. (2017) extended the relations to attributes, verbs, prepositions and pronouns, and performed global inference during test stage. We extend these methods by exploiting the language structure to get context-aware cross-modal representations and learn the matching between grounding entities and their relations jointly.

3 Problem Setting and Overview

The task of general visual grounding aims to localize a set of object regions in an image, each corresponding to a noun phrase in a sentence description of the image. Formally, given an image $I$ and a description $Q$, we denote a set of noun phrases for grounding as $P = \{p_i\}_{i=1}^{N}$ and their corresponding locations as $B = \{b_i\}_{i=1}^{N}$ where $b_i \in \mathbb{R}^4$ is the
bounding box parameters. Our goal is to predict the set B for a given set P from the input I and Q.

To this end, we adopt a hypothesize-and-match strategy that first generates a set of object proposals $O = \{o_m\}_{m=1}^M$ and then formulates the grounding task as a matching problem, in which we seek to establish a cross-modal correspondence between the phrase set P and the object proposal set O. This matching task, nevertheless, is challenging due to large variations in visual and linguistic features, strong context dependency among the grounding entities and the resulting semantic ambiguities in pairwise matching.

To tackle those issues, we propose a language-guided approach motivated by the following three key observations: First, language prior can be used to generate a graph representation of noun phrases and their relations, which captures the global context dependency more effectively than chain-structured models. In addition, the object proposals generated by detectors typically have a high ratio of false positives, and hence it is difficult to encode visual context for each object. We can exploit language structure to guide proposal pruning and build a better context-aware visual representation. Finally, the derived phrase graph structure also includes the phrase relations, which provide additional constraints in the matching for mitigating ambiguities.

We instantiate these ideas by designing a cross-modal graph network for the visual grounding task, which consists of four main modules: a) a **backbone network** that extracts basic linguistic and visual features; b) a **phrase graph network** defined on a language scene graph built from the description to compute the context-aware phrase representations; c) a **visual graph network** defined on a visual scene graph of object proposals constructed under the guidance of the phrase graph, and encodes context cues for the object representations via message propagation; and d) a **graph similarity network** that predicts a global matching of the two graph representations. The overall model is shown in Fig. 1 and we will describe the details of each module in the following section.

### 4 Cross-modal Graph Network

We now introduce our cross-modal graph matching strategy, including the model design of four network modules and the overall inference pipeline, followed by our two-stage model training procedure.

#### 4.1 Backbone Network

Our first network module is a backbone network that takes as input the image I and description Q, and generates corresponding visual and linguistic features. The backbone network consists of two sub-networks: a convolutional network for generating object proposals and a recurrent network for encoding phrases.

Specifically, we adopt the ResNet-101 (He et al. 2016) as our convolutional network to generate feature map $\Gamma$ with channel dimension of $D_0$. We then apply a Region Proposal Network (RPN) (Ren et al. 2015) to generate an initial set of object proposals $O = \{o_m\}_{m=1}^M$, where $o_m \in \mathbb{R}^D$ denotes object location (i.e. bounding box parameters). For each $o_m \in O$, we use RoI-Align (He et al. 2017) and average pooling to compute a feature vector $x^o_{om} \in \mathbb{R}^{D_0}$. We also encode the relative locations of conv-features as a spatial feature vector $x^s_{om}$ (See Suppl. for details), which is fused with $x^o_{om}$ to produce the object representation:

$$x_{om} = F_{vf}([x^o_{om} ; x^s_{om}])$$  \hspace{1cm} (1)

where $x_{om} \in \mathbb{R}^D$, $F_{vf}$ is a multilayer network with fully connected layers and $[;]$ is the concatenate operation.
For the language features, we generate an embedding of noun phrase \( p_i \in \mathcal{P} \). To this end, we first encode each word in sentence \( Q \) into a sequence of word embedding \( \{h_t\}_{t=1:T} \) with a Bi-directional GRU (Chung et al. 2014), where \( T \) is the number of words in sentence. We then compute the phrase representation \( x_{p_i} \), by taking average pooling on the word representations in each \( p_i \):

\[
[ h_1, h_2, \ldots , h_T ] = \text{BiGRU}_p(Q) \tag{2}
\]

\[
x_{p_i} = \frac{1}{|p_i|} \sum_{t \in p_i} h_t \quad i = 1, \ldots , N \tag{3}
\]

where BiGRU\(_p\) denotes the bi-directional GRU, \( h_t, x_{p_i} \in \mathbb{R}^D \) and \( h_t = [h_t^+, h_t^-] \) is the concatenation of forward and backward hidden states for the \( t \)-th word in the sentence.

### 4.2 Phrase Graph Network

To encode the context dependency among phrases, we now introduce our second module, the phrase graph network, which refines the initial phase embedding features by incorporating phrase relations cues in the description.

**Phrase Graph Construction** Specifically, we first build a language scene graph from the image description by adopting an off-the-shelf scene graph parser\(^1\), which also extracts the phrase relations \( \mathcal{R} = \{r_{ij}\} \) from \( Q \), where \( r_{ij} \) is a relationship phrase that connects \( p_i \) and \( p_j \). We denote the language scene graph as \( \mathcal{G}_L = (\mathcal{P}, \mathcal{R}) \) where \( \mathcal{P} \) and \( \mathcal{R} \) are the nodes and edges set respectively. Similar to the phrases in Sec. 4.1, we compute an embedding \( x_{r_{ij}} \) for \( r_{ij} \in \mathcal{R} \) based on a second bi-directional GRU, denoted as BiGRU.

On top of the language scene graph, we construct a phrase graph network that refines the linguistic features through message propagation. Concretly, we associate each node \( p_i \) in the graph \( \mathcal{G}_L \) with its embedding \( x_{p_i} \), and each edge \( r_{ij} \) with its vector representation \( x_{r_{ij}} \). We then define a set of message propagation operators on the graph to generate context-aware representations for all the nodes and edges as follows.

**Phrase Feature Refinement** We introduce two types of message propagation operators to update the node and edge feature respectively. First, to enrich each phrase relation with its subject and object nodes, we send out messages from the noun phrases, which are encoded by their features, to update the relation representation via aggregation:

\[
x_{r_{ij}}^e = x_{r_{ij}} + F_e^l([x_{p_i}; x_{p_j}; x_{r_{ij}}]) \tag{4}
\]

where \( x_{r_{ij}}^e \in \mathbb{R}^D \) is the context-aware relation feature, and \( F_e^l \) is a multilayer network with fully connected layers. The second message propagation operator update each phrase node \( p_i \) by aggregating features from all its neighbour nodes \( \mathcal{N}(i) \) and edges via an attention mechanism:

\[
x_{p_i}^e = x_{p_i} + \sum_{j \in \mathcal{N}(i)} w_{p_j} F_p^l([x_{p_j}; x_{r_{ij}}]) \tag{5}
\]

\( \text{where } x_{p_i}^e \) is the context-aware phrase feature, \( F_p^l \) is a multilayer network, and \( w_{p_j} \) is an attention weight between node \( p_i \) and \( p_j \), which is defined as follows:

\[
w_{p_j} = \text{Softmax}(F_p^c([x_{p_i}; x_{r_{ij}}])^T F_p^l([x_{p_j}; x_{r_{ij}}])) \tag{6}
\]

Here Softmax is a softmax function to compute normalized attention values.

### 4.3 Visual Object Graph Network

Similar to the language counterpart, we also introduce a visual scene graph to capture the global scene context for each object proposal, and to build our third module, the visual object graph network, which enriches object features with their contexts via message propagation over the visual graph.

**Visual Scene Graph Construction** Instead of using a noisy dense graph (Hu et al. 2019), we propose to construct a visual scene graph relevant to the grounding task by exploiting the knowledge of our phrase graph \( \mathcal{G}_L \). To this end, we first prune the object proposal set to keep the objects relevant to the grounding phrases, and then consider only the pairwise relations induced by the phrase graph.

Specifically, we adopt the method in (Plummer et al. 2015; Rohrbach et al. 2016) to select a small set of high-quality proposals \( \mathcal{O} \), for each phrase \( p_i \). To achieve this, we first compute a similarity score \( \phi_{p,m}^i \) for each phrase-boxes pair \( \{p_i, o_m\} \) and a phrase-specific regression offset \( \delta_{p,m}^i \in \mathbb{R}^4 \) for \( o_m \) based on the noun phrase embedding \( x_{p_i} \), and each object feature \( x_{o_m} \) as follows:

\[
\phi_{p,m}^i = F_{cls}^p(x_{p_i}, x_{o_m}), \quad \delta_{p,m}^i = F_{reg}^p(x_{p_i}, x_{o_m}) \tag{7}
\]

where \( F_{cls}^p \) and \( F_{reg}^p \) are two-layer fully-connected networks which transform the input features as in (Lili et al. 2016). We then select the top \( K (K \ll M) \) for each phrase \( p_i \) based on the similarity score \( \phi_{i,m}^p \) and apply the regression offsets \( \delta_{p,m}^i \) to adjust locations of the selected proposals. We denote the refined proposal set of \( p_i \) as \( \mathcal{O}_i = \{o_{k,i}\}_{k=1}^K \) and all the refined proposals as \( \mathcal{V} = \cup_{j=1}^J \mathcal{O}_i \). For each pair of the object proposals \( \{o_{k,i}, o_{k,j}\} \), we introduce an edge \( u_{ij,k} \) if there is a relation \( r_{ij} \) exists in the phrase relation set \( \mathcal{R} \). Denoting the edge set as \( \mathcal{U} = \{u_{ij,k}\} \), we define our visual scene graph as \( \mathcal{G}_V = (\mathcal{V}, \mathcal{U}) \).

Built on top of the visual scene graph, we introduce a visual object graph network that augments the object features with their context through message propagation. Concretely, as in Sec. 4.1, we extract an object feature \( x_{o_{ik}} \) for each proposal \( o_{i,k} \) in \( \mathcal{V} \). Additionally, for each edge \( u_{ij,k} \) in the graph \( \mathcal{G}_V \), we take a union box region of two object \( o_{i,k} \) and \( o_{j,l} \), which is the minimum box region covering both objects, and compute its visual relation feature \( x_{o_{ij,k,l}} \). To do this, we extract a convolution feature \( x_{o_{ij,k,l}}^e \) from \( \mathcal{G}_V \) by Rol-Align, and as in the object features, fuse it with a geometric feature \( x_{o_{ij,k,l}}^g \), encoding location of two objects (See Suppl. for details). We then develop a set of message propagation operators on the graph to generate context-aware representations for all the nodes and edges in the following.

---

\(^1\)https://github.com/vacancy/SceneGraphParser. We refine the language scene graph for the visual grounding task by rule-based post-processing and more details are included in Suppl.
Visual Feature Refinement Similar to Sec. 4.2, we introduce two types of message propagation operators to refine the object and relation features respectively. Specifically, we first update relation features by fusing with their subject and object node features:

\[ x_{u_{ij,k}}^c = x_{u_{ij,k}} + F_r^c([x_{o_{ij}}, x_{o_{ij}}; x_{u_{ij,k}}]) \]  

(8)

where \( F_r^c \) is a multilayer network with fully connected layers. The second type of message update each object node \( o_{ij,k} \) by aggregating features from all its neighbour nodes and corresponding edges via the same attention mechanism:

\[ x_{o_{ij,k}}^c = x_{o_{ij,k}} + \sum_{j,l} \alpha_{ij,kl} F_o^c([x_{o_{ij,l}}; x_{u_{ij,k}}]) \]  

(9)

\[ \alpha_{ij,kl} = \text{Softmax}(F_o^c([x_{o_{ij,l}}; x_{u_{ij,k}}])^T F_o^c([x_{o_{ij,l}}; x_{u_{ij,k}}])) \]

where \( x_{o_{ij,k}}^c \) is the context-aware object feature, \( F_o^c \) is a multilayer network and \( \alpha_{ij,kl} \) is the attention weight between object \( o_{ij,k} \) and \( o_{ij,l} \).

4.4 Graph Similarity Network

Given the phrase and visual scene graph, we formulate the visual grounding as a graph matching problem between two graphs. To solve this, we introduce a graph similarity network to predict the node and edge similarities between the two graphs, followed by a global inference procedure to predict the matching assignment.

Formally, we introduce a similarity score \( \phi_{ij,k} \) for each noun phrase and visual object pair \((x_p^c, x_o^c)\), and an edge similarity score \( \phi_{ij,kl} \) for each phrase and visual relation pair \((x_{r_{ij}}^c, x_{u_{ij,k}}^c)\). For the node similarity \( \phi_{ij,k} \), we first predict a similarity between the refined features \((x_{p_i}^c, x_{o_{ij,k}}^c)\) as in Sec. 4.3, using two-layer fully-connected networks to compute the similarity score and the object offset as follows:

\[ \phi_{ij,k}^o = F_{e}^o(x_{p_i}^c, x_{o_{ij,k}}^{c}) \]  

(10)

\[ \phi_{ij,k}^r = F_{e}^r(x_{r_{ij}}^c, x_{u_{ij,k}}^{c}) \]

We then fuse this with the score used in object pruning to generate the node similarity: \( \phi_{ij,k} = \phi_{ij,k}^o \phi_{ij,k}^r \). The predicted offset is applied to the proposals in the prediction outcome. For the edge similarity, we take the same method as in the node similarity prediction, using a multilayer network \( F_{e}^r \) to predict the edge similarity score \( \phi_{ij,kl} \):

\[ \phi_{ij,kl} = F_{e}^r(x_{r_{ij}}^c, x_{u_{ij,k}}^{c}) \]  

(11)

Given the node and edge similarity scores, we now assign each phrase-object pair a binary variable \( s_{ij,k} \in \{0,1\} \) indicating whether \( o_{ij,k} \) is the target location of \( p_i \). Assuming only one proposal is selected, i.e., \( \sum_{k=1}^{K} s_{ij,k} = 1 \), our subgraph matching can be formulated as a structured prediction problem as follows:

\[ s^* = \arg \max_s \left\{ \sum_{i,k} \phi_{ij,k}s_{ij,k} + \beta \sum_{i,j,k,l} \phi_{ij,kl}s_{ij,k} \cdot s_{ij,l} \right\} \]  

(12)

\[ s.t. \sum_{k=1}^{K} s_{ij,k} = 1; \quad i = 1, \ldots, N \]

where \( \beta \) is a weight balancing the phrase and relation scores. We solve the assignment problem by an approximate algorithm based on exhaustive search with a maximal depth (see Suppl. for detail).

4.5 Model Learning

We adopt a pre-trained ResNet-101 network and an off-the-shelf RPN in our backbone network, and train the remaining network modules. In order to build the visual scene graph, we adopt a two-stage strategy in our model learning. The first stage learns the phrase graph network and object features by a phrase-object matching loss and a box regression loss. We use the learned sub-modules to select a subset of proposals and construct the rest of our model. The second stage trains the entire deep model jointly with a graph similarity loss and a box regression loss.

Specifically, for a noun phrase \( p_i \), the ground-truth for matching scores \( \phi_i^p = \{\phi_{ij,m}^p\}_{m=1}^M \) and \( \phi_i^r = \{\phi_{ij,k}^r\}_{k=1}^K \) are defined as soft label distributions \( Y_i^p = \{y_{ij,m}^p\}_{m=1}^M \) and \( Y_i^r = \{y_{ij,k}^r\}_{k=1}^K \) respectively, based on the IoU between proposal bounding boxes and their ground-truth (Yu et al. 2018b).

Similarly, we compute the ground-truth offset \( \delta_{i,m}^p \) between \( b_i \) and \( o_{ij,m} \), and \( \delta_{i,k}^r \) between \( b_i \) and \( o_{ij,k} \). In addition, the ground-truth for matching scores \( \phi_{ij} = \{\phi_{ij,k}\}_{k=1}^K \) are defined as \( Y_{ij}^r = \{y_{ij,k}\}_{k=1}^K \) based on the IoU between a pair of object proposals \((o_{ij,k}, o_{ij,l})\) and their ground-truth locations \((b_i, b_j)\) (Yang et al. 2018).

After normalizing \( Y_i^p, Y_i^r \) and \( Y_{ij} \) to probability distributions, we define the matching loss \( L_{mat} \) and regression loss \( L_{reg} \) in the first stage as follows:

\[ L_{mat} = \sum_i L_{ce}(\phi_{ij}^p, Y_{ij}^p) \]

\[ L_{reg} = \sum_i \frac{1}{||Y_{ij}^p||_0} \sum_m \|y_{ij,m}^p > 0\| L_{sm}(\delta_{i,m}^p, \delta_{i,m}^p) \]  

(13)

where \( L_{ce} \) is the Cross Entropy loss and \( L_{sm} \) is the Smooth-L1 loss.

For the second stage, the node matching loss \( L_{mat}^c \), edge matching loss \( L_{mat}^e \) and regression loss \( L_{reg}^e \) are defined as:

\[ L_{mat}^c = \sum_i L_{ce}(\phi_{ij}^c, Y_{ij}^c); \quad L_{mat}^e = \sum_{i,j} L_{ce}(\phi_{ij}, Y_{ij}) \]

\[ L_{reg}^c = \sum_i \frac{1}{||Y_{ij}^p||_0} \sum_k \|y_{ij,k}^c > 0\| L_{sm}(\delta_{i,k}^c, \delta_{i,k}^c) \]  

(14)

Where \( ||*||_0 \) is the L0 norm and \( \| \) is the indicator function. Finally the total loss \( L \) can be defined as:

\[ L = L_{mat} + \lambda_1 \cdot L_{reg} + \lambda_2 \cdot L_{mat}^c + \lambda_3 \cdot L_{mat}^e + \lambda_4 \cdot L_{reg}^e \]  

(15)

where \( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \) are weighting coefficients for balancing loss terms.

5 Experiments

5.1 Datasets and Metrics

We evaluate our approach on Flickr30K Entities (Plummer et al. 2015) dataset, which contains 32k images, 275k bounding boxes, and 360k noun phrases. Each image is associated with five sentences description and the noun phrases
Table 1: Results Comparison on Flickr30k test set.

| Methods          | Accuracy(%) |
|------------------|-------------|
| SMPL (Wang et al. 2016) | 42.08       |
| NonlinearSP (Wang, Li, and Lazebnik 2016) | 43.89       |
| GroundR (Rohrbach et al. 2016) | 47.81       |
| MCB (Fukui et al. 2016) | 48.69       |
| RPN (Plummer et al. 2015) | 50.89       |
| Similarity Network (Wang et al. 2018a) | 51.05       |
| IGOP (Yeh et al. 2017) | 53.97       |
| SPC+PPC (Plummer et al. 2017) | 55.49       |
| SS+QRN (Chen, Kossyv, and Nevatia 2017) | 55.99       |
| CITE (Plummer et al. 2018) | 59.27       |
| SeqGROUND (Pelín, Leond, and Markus 2019) | 61.60       |
| Our approach (ResNet-50) | 67.90       |
| DDPPN (Yu et al. 2018b) | 73.30       |
| Our approach (ResNet-101) | 76.74       |

Table 2: Comparison of phrases grounding accuracy over coarse categories on Flickr30k test set.

| Methods          | people | clothing | bodyparts | animal | vehicles | motorcycles | scenic | other |
|------------------|--------|----------|-----------|--------|----------|-------------|--------|-------|
| SMPL             | 57.89  | 34.61    | 15.87     | 55.98  | 52.25    | 32.25       | 25.46  | 31.22 |
| GroundR          | 61.00  | 38.12    | 10.33     | 62.55  | 68.75    | 36.42       | 59.08  | 39.04 |
| IGOE             | 64.73  | 46.68    | 17.21     | 63.83  | 68.72    | 37.65       | 51.39  | 31.77 |
| IGOP             | 68.17  | 56.83    | 19.50     | 70.07  | 73.72    | 39.50       | 60.38  | 32.45 |
| DDPNP            | 72.04  | 55.56    | 25.24     | 76.23  | 65.60    | 35.80       | 51.51  | 35.98 |
| SeqGROUND        | 75.20  | 52.34    | 26.09     | 76.25  | 75.75    | 48.15       | 59.64  | 42.83 |
| Ours (RN-101)    | 83.06  | 63.35    | 9.14      | 84.94  | 76.54    | 55.56       | 51.91  | 52.35 |
| Ours (RN-101)    | 86.82  | 79.92    | 53.54     | 90.73  | 84.75    | 63.58       | 77.12  | 58.65 |

Our model with ResNet-101 backbone achieves 76.74% in accuracy, which improves 3.44% compared to DDPNP (Yu et al. 2018b). For the setting that uses ResNet-50 backbone and a pretrained RPN on MSCOCO (Lin et al. 2014) dataset, we can see that our model achieves 67.90% in accuracy and outperforms SeqGROUND by 6.3%. We also show detailed comparisons per coarse categories in Tab. 2 and it is evident that our approach achieves better performances consistently on most categories.

5.4 Ablation Studies
In this section, we perform several experiments to evaluate the effectiveness of individual components, investigate hyper-parameter $K$ and the impact of relations feature in two graphs in our framework with ResNet-101 as the backbone on Flickr30k val set\(^4\), which is shown in Tab. 3 and Tab. 4.

Baseline: The baseline first predicts the similarity score and regression offset for each phrase-box pair \((x_p, x_b)\), and then selects the most relevant proposal followed by applying its offset. Our baseline grounding accuracy achieves 73.46% with ResNet-101 backbone.

Phrase Graph Net (PGN): PGN propagate language context cues via the scene graph structure effectively. The noun phrases feature can not only be aware of long-term semantic contexts from the other phrases but also enriched by its relation phrases representation. The experiment shows that our PGN can improve the accuracy from 73.46% to 74.40%.

Proposal Pruning (PP): The quality of proposals generation plays an important role in visual grounding task. Here we take proposal pruning operation by utilizing PGN, which can help reduce more ambiguous object candidates with language contexts. We can see a significant improvement of 1.1% accuracy.

Visual Object Graph Net (VOGN): When integrating the VOGN into the whole framework, we can achieve 75.85% accuracy, which is better than the direct matching with the phrase graph. This suggests that the object representation can be more discriminative after conducting message passing among context visual object features\(^5\).

Structural Prediction (SP): The aforementioned PGN and VOGN take the context cues into consideration during their nodes matching. Our approach, by contrast, explicitly

\(^{4}\)We include ablations of ResNet-50 backbone in Suppl.

\(^{5}\)See Suppl. for more experiments that analyze the VOGN.
Table 3: Ablation study on Flickr30K val set.

| Components                  | PGN | PP | VOGN | SP | Acc(%) |
|-----------------------------|-----|----|------|----|--------|
| Baseline                    | ✓   | -  | -    | -  | 74.40  |
| ✓                           | ✓   | -  | -    | -  | 75.50  |
| ✓                           | ✓   | ✓  | -    | -  | 75.85  |
| ✓                           | ✓   | ✓  | ✓    | ✓  | 76.19  |
| Ours                        | ✓   | ✓  | ✓    | ✓  |        |

Table 4: Ablation study of K proposals on Flickr30K val set.

| K  | Acc(%) |
|----|--------|
| 5  | 74.97  |
| 10 | 76.19  |
| 20 | 76.07  |

Figure 2: Visualization of phrase grounding results in Flickr30K val set. The colored bounding boxes, which are predicted by our approach, correspond to the noun phrases in the sentences with the same color. The dot boxes denote the predicted results without relations constraint, while the white boxes are ground-truths and the red boxes are the incorrect predictions. The last column is the failure cases.

5.5 Qualitative Visualization Results

We show some qualitative visual grounding results in Fig.2 to demonstrate the capabilities of our framework in challenging scenarios. In (a1) and (a2), our framework is able to successfully localize multiple entities in the long sentences without ambiguity. With the help of VOGN, we can see that our model localize a mug close to man correctly rather than another mug in the left bottom in (b1). Column 3 shows that relations constraint can help refine the final prediction. The last column is failure cases. Our model cannot ground objects in images correctly with severe visual ambiguity.

6 Conclusion

In this paper, we have proposed a context-aware cross-modal graph network for visual grounding task. Our method exploits a graph representation for language description, and transfers the linguistic structure to object proposals to build a visual scene graph. Then we use message propagation to extract global context representations both for the grounding entities and visual objects. As a result, it is able to conduct a global matching between both graph nodes and relation edges. We present a modular graph network to instantiate our core idea of context-aware cross-modal matching. Moreover, we adopt a two-stage strategy in our model learning, of which the first stage learns a phrase graph network and visual object features while the second stage trains the entire deep network jointly. Finally, we achieve the state-of-the-art performances on Flickr30K Entities benchmark, and outperform other approaches by a sizable margin.
References

[Anderson et al. 2018] Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In CVPR.

[Cadene et al. 2019] Cadene, R.; Ben-Younes, H.; Thome, N.; and Cord, M. 2019. Murel: Multimodal Relational Reasoning for Visual Question Answering. In CVPR.

[Chen, Kovuvi, and Nevatia 2017] Chen, K.; Kovuvi, R.; and Nevatia, R. 2017. Query-guided regression network with context policy for phrase grounding. In ICCV.

[Chung et al. 2014] Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

[Das et al. 2017] Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2017. Visual dialog. In CVPR.

[Devlin et al. 2018] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[Feng et al. 2019] Feng, Y.; Ma, L.; Liu, W.; and Luo, J. 2019. Unsupervised image captioning. In CVPR.

[Fukui et al. 2016] Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In CVPR.

[He et al. 2017] He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In ICCV.

[Hu et al. 2019] Hu, R.; Rohrbach, A.; Darrell, T.; and Saenko, K. 2019. Language-conditioned graph networks for relational reasoning. arXiv preprint arXiv:1905.04405.

[Justin et al. 2015] Justin, J.; Ranjay, K.; Michael, S.; Li-Jia, L.; David, A. S.; and Li, F.-F. 2015. Image retrieval using scene graphs. In CVPR.

[Kottur et al. 2018] Kottur, S.; Moura, J. M. F.; Parikh, D.; Batra, D.; and Rohrbach, M. 2018. Visual coreference resolution in visual dialog using neural module networks. In ECCV.

[Krishna et al. 2017] Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV.

[Li et al. 2017] Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene graph generation from objects, phrases and region captions. In ICCV.

[Lili et al. 2016] Lili, M.; Rui, M.; Ge, L.; Yan, X.; Lu, Z.; Rui, Y.; and Zhi, J. 2016. Natural language inference by tree-based convolution and heuristic matching. In ACL.

[Lin et al. 2014] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In ECCV.

[Mogadala, Kalimuthu, and Klakow 2019] Mogadala, A.; Kalimuthu, M.; and Klakow, D. 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. arXiv preprint arXiv:1907.09358.

[Mun et al. 2018] Mun, J.; Lee, K.; Shin, J.; and Han, B. 2018. Learning to specialize with knowledge distillation for visual question answering. In NeurIPS.

[Nagaraja, Morariu, and Davis 2016] Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling context between objects for referring expression understanding. In ECCV.

[Nam et al. 2019] Nam, V.; Lu, J.; Chen, S.; Kevin, M.; Li-Jia, L.; Li, F.-F.; and James, H. 2019. Composing text and image for image retrieval - an empirical odyssey. In CVPR.

[Pelin, Leonid, and Markus 2019] Pelin, D.; Leonid, S.; and Markus, G. 2019. Neural sequential phrase grounding (seq-ground). In CVPR.

[Peng et al. 2019] Peng, W.; Qi, W.; Jiewei, C.; Chunhua, S.; Lianli, G.; and Anton, v. d. H. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In CVPR.

[Peters et al. 2018] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, E.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In NAACL.

[Plummer et al. 2015] Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In ICCV.

[Plummer et al. 2017] Plummer, B. A.; Mallya, A.; Cervantes, C. M.; Hockenmaier, J.; and Lazebnik, S. 2017. Phrase localization and visual relationship detection with comprehensive image-language cues. In ICCV.

[Plummer et al. 2018] Plummer, B. A.; Kordas, P.; Hadi Kiapour, M.; Zheng, S.; Piramutlu, R.; and Lazebnik, S. 2018. Conditional image-text embedding networks. In ECCV.

[Ren et al. 2015] Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In NeurIPS.

[Rohrbach et al. 2016] Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of textual phrases in images by reconstruction.

[Schuster et al. 2015] Schuster, S.; Krishna, R.; Chang, A.; Fei-Fei, L.; and Manning, C. D. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In Workshop on Vision and Language (VL15). Association for Computational Linguistics.

[Wang et al. 2016] Wang, M.; Azab, M.; Kojima, N.; Mihalcea, R.; and Deng, J. 2016. Structured matching for phrase localization. In ECCV.

[Wang et al. 2018a] Wang, L.; Li, Y.; Huang, J.; and Lazebnik, S. 2018a. Learning two-branch neural networks for image-text matching tasks. IEEE TPAMI.

[Wang et al. 2018b] Wang, Y.-S.; Liu, C.; Zeng, X.; and Yuille, A. 2018b. Scene graph parsing as dependency Parsing. In NAACL.

[Wang, Li, and Lazebnik 2016] Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. In CVPR.

[Yang et al. 2018] Yang, J.; Lu, J.; Lee, S.; Batra, D.; and Parikh, D. 2018. Graph r-cnn for scene graph generation. In ECCV.

[Yeh et al. 2017] Yeh, R.; Xiong, J.; Hwu, W.-M.; Do, M.; and Schwing, A. 2017. Interpretable and globally optimal prediction for textual grounding using image concepts. In NeurIPS.

[Yu et al. 2018a] Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018a. Mattnet: Modular attention network for referring expression comprehension. In CVPR.

[Yu et al. 2018b] Yu, Z.; Yu, J.; Xiang, C.; Zhao, Z.; Tian, Q.; and Tao, D. 2018b. Rethinking diversified and discriminative proposal generation for visual grounding. arXiv preprint arXiv:1805.03508.
[Zhang et al. 2017] Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017. Visual translation embedding network for visual relation detection. In CVPR.
Learning Cross-modal Context Graph Networks for Visual Grounding

Yongfei Liu1∗ Bo Wan1∗ Xiaodan Zhu2 Xuming He1
1 ShanghaiTech University 2 Queens University 
{liuyf3, wanbo, hexm}@shanghaitech.edu.cn xiaodan.zhu@queensu.ca

1 Cross-modal Graph Network

1.1 Spatial Feature of Object
We generate a coordinate map $\alpha$ with the same spatial size as the convolution feature map $\Gamma$. The coordinate map $\alpha$ consists of two channels, indicating the x, y coordinates for each pixel in $\Gamma$, and normalized by the feature map center. For each object proposal $o_m \in \mathbb{R}^4$, we crop a coordinate map from $\alpha$ with RoI-Align and embed it into a spatial feature vector $x_{so} \in \mathbb{R}^{256}$ by multiple fully connection layers.

1.2 Spatial Feature of Union Region
We generate a two-channel binary mask for $o_{i,k}$ and $o_{j,l}$ separately where locations within object proposal $o_{i,k}$, $o_{j,l}$ fill 1 and others fill 0. Then the two-channel binary mask is resized to $64 \times 64$. And we use multiple fully connected layers to embed it to a geometric feature vector $x_{su} \in \mathbb{R}^{256}$.

1.3 Scene Graph Parser
For a given sentence, we use a public toolkit1 to generate a language scene graph, in which nodes encode noun phrases and edges are the relationships between them. In this language scene graph parser, a dependency parser is first applied to the input sentence and then hand-craft rules are employed to generate language scene graphs. However, we observe some issues associated with the off-the-shelf parser: 1) noun phrases in the parses sometimes do not correspond to the given phrases; 2) some phrases and their relationships are still missing the in parses.

To address the aforementioned limitations, we perform additional post-processing on the Flickr30K Entities dataset. First, we take all given phrases as graph nodes. For each phrase, we pick a noun phrase in the parse that has a maximum word overlap with this given phrase. We then assign the parsed relations to these nodes. However, there are still some isolated nodes in the resulting graph. We further recall some missing relations by taking advantage of the coarse categories of the given phrases. Specifically, for an isolated phrase, if its type is clothing or bodyparts, we find a phrase with the type of people as its subject, and assign a relationship wear / have to them. If there are multiple phrases with the type of people in the graph nodes, we select the one that has a minimum word distance in the sentence with the isolated phrase. The motivation of our rules design comes from the observation that most of clothing / bodyparts phrases are related to a people phrase, and their relationships are generally wear / have.

1.4 Solving Structured Prediction
We solve the structured prediction problem by taking an exhaustive search on all the possibilities of $s$ in Equ. 12 with a maximal depth when noun phrase number $N$ is less than 6, and applying only node matching between the phrase

Figure 1: Illustration of spatial feature embedding.

Figure 2: Illustration of pairwise geometric feature embedding.
The motivation of the solving strategy comes from the observation that 96.12% language scene graphs in Flickr30K dataset have less than 6 nodes. The complexity of exhaustive search with a maximal depth is $K^N$, which is not time-consuming when $N$ is small.

2 Experiments

2.1 Model details with ResNet-50 backbone

We take an off-the-shelf object detector with ResNet-50 as its backbone to generate the initial set of proposals. It is based on FasterRCNN and pre-trained on the MSCOCO dataset (Lin et al. 2014). Other settings are same to the model with ResNet-10 backbone. During the training stage, we use SGD optimizer with initial learning rate $1e^{-1}$, weight decay $1e^{-4}$ and momentum 0.9. The model is trained with 60k iterations totally with batch size 24, and decay the learning rate 10 times in 20k and 40k iterations respectively.

2.2 Ablations with ResNet-50 backbone

Table 1: Ablation studies on Flickr30K val set with ResNet-50 backbone.

| Components | Methods | PGN | PP | VOGN | SP | Acc(%) |
|------------|---------|-----|----|------|----|--------|
| Baseline   |         | -   | -  | -    | -  | 60.31  |
| ✓          |         | -   | -  | -    | -  | 62.45  |
| ✓          |         | ✓   | -  | -    | -  | 67.51  |
| ✓          |         | ✓   | ✓  | -    | -  | 67.77  |
| Ours       |         | ✓   | ✓  | ✓    | ✓  | 68.12  |

In order to investigate the effectiveness of the individual component of our framework with ResNet-50 backbone, we also conduct a series of ablation studies. As shown in Tab. 1, the accuracy shows the same growth trend compared to ResNet-101 backbone. In particular, we can observe a significant performance improvement when adopting proposal pruning over baseline model, which improves the accuracy from 60.31% to 66.77%. This indicates that proposal pruning is critical for visual grounding task when the object detector doesn’t perform well.

2.3 Additional Experiments on VOGN

To validate the effectiveness of VOGN, we conduct some additional experiments as shown in Tab. 2. In the baseline model, we compute the similarity score and regression offset for each phrase-box pair $(x_{p_i}, x_{o_m})$. Then we adopt proposal pruning strategy over baseline model without PGN, which can improve grounding accuracy from 73.46% to 74.6%.

Table 2: Additional Experiments of VOGN with ResNet-101 backbone on Flickr30K val set.

| Components | Methods | PGN | PP | VOGN | SP | Acc(%) |
|------------|---------|-----|----|------|----|--------|
| Baseline   |         | -   | -  | -    | -  | 73.46  |
| -          |         | ✓   | -  | -    | -  | 74.60  |
| -          |         | ✓   | ✓  | -    | -  | 75.59  |
| -          |         | ✓   | ✓  | ✓    | ✓  | 74.80  |

Furthermore, we add our VOGN under this setting and observe a significant improvement from 74.60% to 75.59%, which indicates the visual object representation can be more discriminative with its context cues.

Finally, the performance will drop sharply from 75.59% to 74.80% without considering visual relations feature $x_{c_{ij,k}}$ during message passing, which suggests that visual relations play an important role in computing attention among objects.