Learning Cross-Modal Context Graph for Visual Grounding

Yongfei Liu,[1]* Bo Wan,[1]* Xiaodan Zhu,[2] Xuming He[1]
[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).

General visual grounding typically attempts to localize object regions that correspond to *multiple* noun phrases in image descriptions. Despite significant progress in solving vision (Ren et al. 2015; Zhang et al. 2017) or language (Peters et al. 2018; Devlin et al. 2018) tasks, it remains challenging to establish such cross-modal correspondence between objects and phrases, mainly because of large variations in object appearances and phrase descriptions, strong context dependency among these grounding entities, and the resulting semantic ambiguities in their representations (Plummer et al. 2015; 2018).

Many existing works on visual grounding tackle the problem by localizing each noun phrase independently via phrase-object matching (Plummer et al. 2015; 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-

*Both authors contributed equally to the work. This work was supported by Shanghai NSF Grant (No. 18ZR1425100) and NSFC Grant (No. 61703195). The research of 3rd author is supported by the Discovery Grant of Natural Sciences and Engineering Research Council of Canada.

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 $\mathcal{P} = \{p_i\}_{i=1}^{N}$ and their corresponding locations as $\mathcal{B} = \{b_i\}_{i=1}^{N}$ where $b_i \in \mathbb{R}^4$ is the

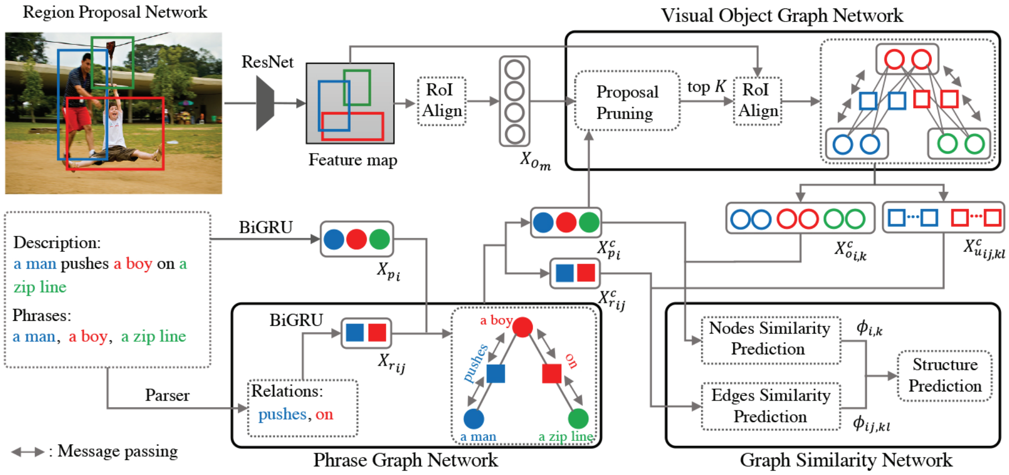


Figure 1: Model Overview: There are four modules in our network, the **Backbone Network** extracts basic linguistic and visual features; the **Phrase Graph Network** is defined on the a parsed language scene graph to refine language representations; the **Visual Object Graph Network** is defined on a visual scene graph which is constructed under the guidance of the phrase graph to refine visual object feature; finally a **Graph Similarity Network** predicts the global matching of those two graph representations. ***Solid circles*** denote noun phrase features while ***solid squares*** represent relation phrase features. ***Hollow circles and squares*** denote visual object and relation features respectively.

bounding box parameters. Our goal is to predict the set $\mathcal{B}$ for a given set $\mathcal{P}$ from the input I and Q.

To this end, we adopt a hypothesize-and-match strategy that first generates a set of object proposals $\mathcal{O} = \{o_m\}_{m=1}^M$ and then formulates the grounding task as a matching problem, in which we seek to establish a cross-modal correspondence between the phrase set $\mathcal{P}$ and the object proposal set $\mathcal{O}$. 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 $\mathbf{\Gamma}$ 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 $\mathcal{O} = \{o_m\}_{m=1}^M$, where $o_m \in \mathbb{R}^4$ denotes object location (i.e. bounding box parameters). For each $o_m \in \mathcal{O}$, we use RoI-Align (He et al. 2017) and average pooling to compute a feature vector $\mathbf{x}^a_{o_m} \in \mathbb{R}^{D_0}$. We also encode the relative locations of conv-features as a spatial feature vector $\mathbf{x}^s_{o_m}$ (See Suppl. for details), which is fused with $\mathbf{x}^a_{o_m}$ to produce the object representation:

$$\mathbf{x}_{o_m} = F_{vf}([\mathbf{x}^a_{o_m}; \mathbf{x}^s_{o_m}]) \tag{1}$$

where $\mathbf{x}_{o_m} \in \mathbb{R}^D$, 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 $\{h_t\}_{t=1 \ldots T}$ with a Bi-directional GRU (Chung et al. 2014), where T is the number of words in sentence. We then compute the phrase representation $\mathbf{x}_{p_i}$ by taking average pooling on the word representations in each p_i:

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

$$\mathbf{x}_{p_i} = \frac{1}{|p_i|} \sum_{t \in p_i} h_t \quad i = 1, \cdots, N \tag{3}$$

where BiGRU$_p$ denotes the bi-directional GRU, $h_t, \mathbf{x}_{p_i} \in \mathbb{R}^D$ and $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$ is the concatenation of forward and backward hidden states for 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 $\mathbf{x}_{r_{ij}}$ for $r_{ij} \in \mathcal{R}$ based on a second bi-directional GRU, denoted as BiGRU$_r$.

On top of the language scene graph, we construct a phrase graph network that refines the linguistic features through message propagation. Concretely, we associate each node p_i in the graph $\mathcal{G}_L$ with its embedding $\mathbf{x}_{p_i}$, and each edge r_{ij} with its vector representation $\mathbf{x}_{r_{ij}}$. 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:

$$\mathbf{x}^c_{r_{ij}} = \mathbf{x}_{r_{ij}} + F^l_e([\mathbf{x}_{p_i}; \mathbf{x}_{p_j}; \mathbf{x}_{r_{ij}}]) \tag{4}$$

where $\mathbf{x}^c_{r_{ij}} \in \mathbb{R}^D$ is the context-aware relation feature, and F^l_e 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:

$$\mathbf{x}^c_{p_i} = \mathbf{x}_{p_i} + \sum_{j \in \mathcal{N}(i)} w_{p_{ij}} F^l_p([\mathbf{x}_{p_j}; \mathbf{x}^c_{r_{ij}}]) \tag{5}$$

where $\mathbf{x}^c_{p_i}$ is the context-aware phrase feature, F^l_p is a multilayer network, and $w_{p_{ij}}$ is an attention weight between node p_i and p_j, which is defined as follows:

$$w_{p_{ij}} = \underset{j \in \mathcal{N}(i)}{\text{Softmax}}(F^l_p([\mathbf{x}_{p_i}; \mathbf{x}^c_{r_{ij}}])^\intercal F^l_p([\mathbf{x}_{p_j}; \mathbf{x}^c_{r_{ij}}])) \tag{6}$$

Here $\mathrm{Softmax}$ 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 $\mathcal{O}_i$ for each phrase p_i. To achieve this, we first compute a similarity score $\phi^p_{i,m}$ for each phrase-boxes pair $\langle p_i, o_m \rangle$ and a phrase-specific regression offset $\delta^p_{i,m} \in \mathbb{R}^4$ for o_m based on the noun phrase embedding $\mathbf{x}^c_{p_i}$ and each object feature $\mathbf{x}_{o_m}$ as follows:

$$\phi^p_{i,m} = F^p_{cls}(\mathbf{x}^c_{p_i}, \mathbf{x}_{o_m}), \quad \delta^p_{i,m} = F^p_{reg}(\mathbf{x}^c_{p_i}, \mathbf{x}_{o_m}) \tag{7}$$

where F^p_{cls} and F^p_{reg} 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 $\phi^p_{i,m}$, and apply the regression offsets $\delta^p_{i,m}$ to adjust locations of the selected proposals. We denote the refined proposal set of p_i as $\mathcal{O}_i = \{o_{i,k}\}^K_{k=1}$ and all the refined proposals as $\mathcal{V} = \cup^N_{i=1} \mathcal{O}_i$. For each pair of the object proposals $\langle o_{i,k}, o_{j,l} \rangle$, we introduce an edge $u_{ij,kl}$ if there is a relation r_{ij} exists in the phrase relation set $\mathcal{R}$. Denoting the edge set as $\mathcal{U} = \{u_{ij,kl}\}$, 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 $\mathbf{x}_{o_{i,k}}$ for each proposal $o_{i,k}$ in $\mathcal{V}$. Additionally, for each edge $u_{ij,kl}$ 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 $\mathbf{x}_{u_{ij,kl}}$. To do this, we extract a convolution feature $\mathbf{x}^a_{u_{ij,kl}}$ from $\mathbf{\Gamma}$ by RoI-Align, and as in the object features, fuse it with a geometric feature $\mathbf{x}^s_{u_{ij,kl}}$ 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:

$$\mathbf{x}^c_{u_{ij,kl}} = \mathbf{x}_{u_{ij,kl}} + F^v_e([\mathbf{x}_{o_{i,k}}; \mathbf{x}_{o_{j,l}}; \mathbf{x}_{u_{ij,kl}}]) \tag{8}$$

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

$$\mathbf{x}^c_{o_{i,k}} = \mathbf{x}_{o_{i,k}} + \sum_{j,l} \alpha_{ij,kl} F^v_o([\mathbf{x}_{o_{j,l}}; \mathbf{x}^c_{u_{ij,kl}}]) \tag{9}$$

$$\alpha_{ij,kl} = \underset{j,l}{\mathrm{Softmax}}(F^v_o([\mathbf{x}_{o_{i,k}}; \mathbf{x}^c_{u_{ij,kl}}])^\intercal F^v_o([\mathbf{x}_{o_{j,l}}; \mathbf{x}^c_{u_{ij,kl}}])$$

where $\mathbf{x}^c_{o_{i,k}}$ is the context-aware object feature, F^v_o is a multilayer network and $\alpha_{ij,kl}$ is the attention weight between object $o_{i,k}$ and $o_{j,l}$.

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 $\phi_{i,k}$ for each noun phrase and visual object pair $\langle \mathbf{x}^c_{p_i}, \mathbf{x}^c_{o_{i,k}} \rangle$, and an edge similarity score $\phi_{ij,kl}$ for each phrase and visual relation pair $\langle \mathbf{x}^c_{r_{i,j}}, \mathbf{x}^c_{u_{ij,kl}} \rangle$. For the *node similarity* $\phi_{i,k}$, we first predict a similarity between the refined features $\langle \mathbf{x}^c_{p_i}, \mathbf{x}^c_{o_{i,k}} \rangle$ as in Sec. 4.3, using two-layer fully-connected networks to compute the similarity score and the object offset as follows,

$$\phi^g_{i,k} = F^g_{cls}(\mathbf{x}^c_{p_i}, \mathbf{x}^c_{o_{i,k}}) \qquad \delta^g_{i,k} = F^g_{reg}(\mathbf{x}^c_{p_i}, \mathbf{x}^c_{o_{i,k}}) \tag{10}$$

We then fuse this with the score used in object pruning to generate the node similarity: $\phi_{i,k} = \phi^p_{i,k} \cdot \phi^g_{i,k}$. 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^r_{cls} to predict the edge similarity score $\phi_{ij,kl}$:

$$\phi_{ij,kl} = F^r_{cls}(\mathbf{x}^c_{r_{ij}}, \mathbf{x}^c_{u_{ij,kl}}) \tag{11}$$

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

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} \Big\{ \sum_{i,k} \phi_{i,k} s_{i,k} + \beta \sum_{i,j,k,l} \phi_{ij,kl} s_{i,k} \cdot s_{j,l} \Big\}$$

$$s.t. \sum_{k=1}^{K} s_{i,k} = 1; \quad i = 1, \ldots, N \tag{12}$$

where β 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 $\boldsymbol{\phi}^p_i = \{\phi^p_{i,m}\}^M_{m=1}$ and $\boldsymbol{\phi}^g_i = \{\phi^g_{i,k}\}^K_{k=1}$ are defined as soft label distributions $\boldsymbol{Y}^p_i = \{y^p_{i,m}\}^M_{m=1}$ and $\boldsymbol{Y}^g_i = \{y^g_{i,k}\}^K_{k=1}$ respectively, based on the IoU between proposal bounding boxes and their ground-truth (Yu et al. 2018b).

Similarly, we compute the ground-truth offset $\delta^{p*}_{i,m}$ between b_i and o_m, $\delta^{g*}_{i,k}$ between b_i and $o_{i,k}$. In addition, the ground-truth for matching scores $\boldsymbol{\phi}^r_{ij} = \{\phi_{ij,kl}\}^K_{k,l=1}$ are defined as $\boldsymbol{Y}^r_{ij} = \{y^r_{ij,kl}\}^K_{k,l=1}$ based on the IoU between a pair of object proposals $\langle o_{i,k}, o_{j,l} \rangle$ and their ground-truth locations $\langle b_i, b_j \rangle$ (Yang et al. 2018).

After normalizing $\boldsymbol{Y}^p_i$, $\boldsymbol{Y}^g_i$ and $\boldsymbol{Y}^r_{ij}$ to probability distributions, we define the matching loss $\mathcal{L}^p_{mat}$ and regression loss $\mathcal{L}^p_{reg}$ in the first stage as follows:

$$\mathcal{L}^p_{mat} = \sum_i L_{ce}(\boldsymbol{\phi}^p_i, \boldsymbol{Y}^p_i)$$

$$\mathcal{L}^p_{reg} = \sum_i \frac{1}{||\boldsymbol{Y}^p_i||_0} \sum_m \mathbb{I}(y^p_{i,m} > 0) L_{sm}(\delta^p_{i,m}, \delta^{p*}_{i,m}) \tag{13}$$

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

For the second stage, the node matching loss $\mathcal{L}^g_{mat}$, edge matching loss $\mathcal{L}^r_{mat}$ and regression loss $\mathcal{L}^g_{reg}$ are defined as:

$$\mathcal{L}^g_{mat} = \sum_i L_{ce}(\boldsymbol{\phi}^g_i, \boldsymbol{Y}^g_i), \quad \mathcal{L}^r_{mat} = \sum_{i,j} L_{ce}(\boldsymbol{\phi}^r_{ij}, \boldsymbol{Y}^r_{ij})$$

$$\mathcal{L}^g_{reg} = \sum_i \frac{1}{||\boldsymbol{Y}^g_i||_0} \sum_k \mathbb{I}(y^g_{i,k} > 0) L_{sm}(\delta^g_{i,k}, \delta^{g*}_{i,k}) \tag{14}$$

Here $|| * ||_0$ is the L0 norm and $\mathbb{I}$ is the indicator function. Finally the total loss $\mathcal{L}$ can be defined as:

$$\begin{aligned} \mathcal{L} = &\mathcal{L}^p_{mat} + \lambda_1 \cdot \mathcal{L}^p_{reg} \\ &+ \lambda_2 \cdot \mathcal{L}^g_{mat} + \lambda_3 \cdot \mathcal{L}^r_{mat} + \lambda_4 \cdot \mathcal{L}^g_{reg} \end{aligned} \tag{15}$$

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
GroundeR (Rohrbach et al. 2016)	47.81
MCB (Fukui et al. 2016)	48.69
RtP (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, Kovvuri, and Nevatia 2017)	55.99
CITE (Plummer et al. 2018)	59.27
SeqGROUND (Pelin, Leonid, and Markus 2019)	61.60
Our approach (ResNet-50)	**67.90**
DDPN (Yu et al. 2018b)	73.30
Our approach (ResNet-101)	**76.74**

are provided with their corresponding bounding boxes in the image. Following (Rohrbach et al. 2016), if a single noun phrase corresponds to multiple ground-truth bounding boxes, we merge the boxes and use the union region as their ground-truth. We adopt the standard dataset split as in Plummer et al. (2015), which separates the dataset into 30k images for training, 1k for validation and 1k for testing. We consider a noun phrase grounded correctly when its predicted box has at least 0.5 IoU with its ground-truth location. The grounding accuracy (i.e., Recall@1) is the fraction of correctly grounded noun phrases.

5.2 Implementation Details

We generate an initial set of $M = 100$ object proposals with a RPN from Anderson et al. (2018)[2]. We use the output of ResNet C4 block as our feature map $\mathbf{\Gamma}$ with channel dimension $D_0 = 2048$ and the visual object features are obtained by applying RoI-Align with resolution 14×14 on $\mathbf{\Gamma}$. The embedding dimension D of phrase and visual representation is set as 1024. In visual graph construction, we select the most $K = 10$ relevant object candidates for each noun phrase.

For model training, we use SGD optimizer with initial learning rate 5e-2, weight decay 1e-4 and momentum 0.9. We train 60k iterations with batch-size 24 totally and decay the learning rate 10 times in 20k and 40k iterations respectively. The loss weights of regression terms λ_1 and λ_4 are set to 0.1 while matching terms λ_2 and λ_3 are set to 1. During the test stage, we search an optimal weight $\beta^* \in [0, 1]$ on val set and apply it to test set directly.

5.3 Results and Comparisons

We report the performance of the proposed framework on the Flickr30K Entities test set and compare it with several the state-of-the-art approaches. Here we consider two model configurations for proper comparisons, which use an ResNet-50[3] and an ResNet-101 as their backbone network, respectively.

Table 2: Comparison of phrases grounding accuracy over coarse categories on Flickr30K test set.

Methods	people	clothing	bodyparts	animal	vehicles	instruments	scene	other
SMPL	57.89	34.61	15.87	55.98	52.25	23.46	34.22	26.23
GroundR	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	68.17	56.83	19.50	70.07	73.72	39.50	60.38	32.45
SS+QRN	68.24	47.98	20.11	73.94	73.66	29.34	66.00	38.32
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
Ours (RN-50)	**83.06**	**63.35**	24.28	**84.94**	**78.25**	**55.56**	61.67	**52.05**
Ours (RN-101)	**86.82**	**79.92**	**53.54**	**90.73**	**84.75**	**63.58**	**77.12**	**58.65**

As shown in Tab. 1, our approach outperforms the prior methods by a large margin in both settings. In particular, our model with ResNet-101 backbone achieves **76.74%** in accuracy, which improves 3.44% compared to DDPN (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 $\langle \mathbf{x}_{p_i}, \mathbf{x}_{o_m} \rangle$, 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].

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

[2] It is based on FasterRCNN (Ren et al. 2015) with ResNet-101 as its backbone, trained on Visual Genome dataset (Krishna et al. 2017). We use its RPN to generate object proposals.

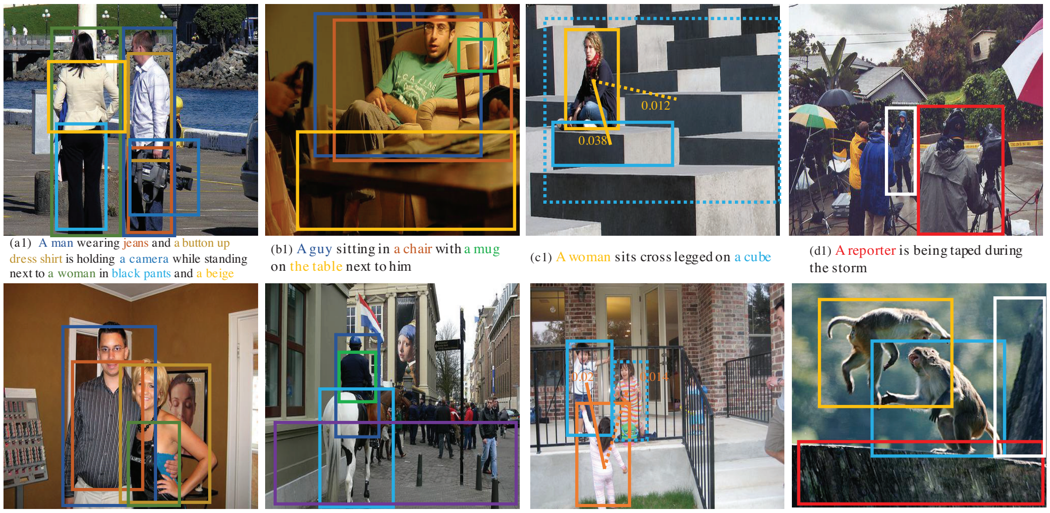
[3] Model details of ResNet-50 backbone are included in Suppl.

[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					**Components (w/o relations feature)**			
Methods	PGN	PP	VOGN	SP	**Acc(%)**	PGN	PP	VOGN	**Acc(%)**
Baseline	-	-	-	-	73.46	-	-	-	-
	✓	-	-	-	74.40	✓(w/o $\mathbf{x}^c_{r_{ij}}$)	-	-	74.11
	✓	✓	-	-	75.50	✓(w/o $\mathbf{x}^c_{r_{ij}}$)	✓	-	75.32
	✓	✓	✓	-	75.85	✓(w/o $\mathbf{x}^c_{r_{ij}}$)	✓	✓(w/o $\mathbf{x}^c_{u_{ij,kl}}$)	75.44
Ours	✓	✓	✓	✓	**76.19**	-	-	-	-



(a1) A man wearing jeans and a button up dress shirt is holding a camera while standing next to a woman in black pants and a beige

(b1) A guy sitting in a chair with a mug on the table next to him

(c1) A woman sits cross legged on a cube

(d1) A reporter is being taped during the storm

(a2) A man in a striped shirt hugging a blond short-haired woman with a black apron on

(b2) A man in a black jacket is riding a horse on a public sidewalk

(c2) The girl is stretching to give the boy the racket

(d2) One monkey is jumping near another monkey on the top of a wooden structure

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.

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

K	5	10	20
Acc(%)	74.97	**76.19**	76.07

takes the cross-modal relation matching into account and predicts the final result via a global optimization. We can see further improvement of accuracy from 75.85% to 76.19%.

Hyper-parameter K and Relations Feature: In Tab.4, our framework achieves the highest accuracy when $K = 10$ while $K = 5$ will result in performance dropping from 76.19% to 74.97% due to the lower proposals recall. When $K = 20$, our model will get a comparable performance but consume more computation resources and inference time.

We also perform experiments to show the impact of relation phrases and visual relations in PGN and VOGN in Tab. 3. For PGN, the performance will drop from 74.40% to 74.11% without phrase relations $\mathbf{x}^c_{r_{ij}}$. And we can see 0.41% performance drop when ignoring both phrase relations $\mathbf{x}^c_{r_{ij}}$ and visual relations $\mathbf{x}^c_{u_{ij,kl}}$ in PGN and VOGN.

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

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

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, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *CVPR*.

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, Y.; Ma, L.; Liu, W.; and Luo, J. 2019. Unsupervised image captioning. In *CVPR*.

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, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*.

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

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, 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, 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, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene graph generation from objects, phrases and region captions. In *ICCV*.

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, 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, 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, J.; Lee, K.; Shin, J.; and Han, B. 2018. Learning to specialize with knowledge distillation for visual question answering. In *NeuIPS*.

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

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, D.; Leonid, S.; and Markus, G. 2019. Neural sequential phrase grounding (seqground). In *CVPR*.

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, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*.

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, 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, B. A.; Kordas, P.; Hadi Kiapour, M.; Zheng, S.; Piramuthu, R.; and Lazebnik, S. 2018. Conditional image-text embedding networks. In *ECCV*.

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

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

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, M.; Azab, M.; Kojima, N.; Mihalcea, R.; and Deng, J. 2016. Structured matching for phrase localization. In *ECCV*.

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

Wang, Y.-S.; Liu, C.; Zeng, X.; and Yuille, A. 2018b. Scene graph parsing as dependency parsing. In *NAACL*.

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

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

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 *NeuIPS*.

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, 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, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017. Visual translation embedding network for visual relation detection. In *CVPR*.