Improving Zero-Shot Phrase Grounding via Reasoning on External Knowledge and Spatial Relations

Zhan Shi[*†],[1] **Yilin Shen,**[2] **Hongxia Jin,**[2] **Xiaodan Zhu**[1]

[1]Ingenuity Labs Research Institute & ECE, Queen's University
[2]Samsung Research America
{z.shi, xiaodan.zhu}@queensu.ca, {yilin.shen, hongxia.jin}@samsung.com

Abstract

Phrase grounding is a multi-modal problem that localizes a particular noun phrase in an image referred to by a text query. In the challenging zero-shot phrase grounding setting, the existing state-of-the-art grounding models have limited capacity in handling the unseen phrases. Humans, however, can ground novel types of objects in images with little effort, significantly benefiting from reasoning with commonsense. In this paper, we design a novel phrase grounding architecture that builds multi-modal knowledge graphs using external knowledge and then performs graph reasoning and spatial relation reasoning to localize the referred nouns phrases. We perform extensive experiments on different zero-shot grounding splits sub-sampled from the Flickr30K Entity and Visual Genome dataset, demonstrating that the proposed framework is orthogonal to backbone image encoders and outperforms the baselines by 2∼3% in accuracy, resulting in a significant improvement under the standard evaluation metrics.

Introduction

Localizing objects in an image referenced by noun phrases in a query (Chen, Kovvuri, and Nevatia 2017; Plummer et al. 2015; Yu et al. 2018), a fundamental problem known as *phrase grounding* or *referring expressions*, has drawn extensive attention in both the natural language processing and computer vision community. A good phrase grounding system can benefit many other downstream tasks such as visual question answering (Antol et al. 2015; Goyal et al. 2017), image retrieval (Johnson et al. 2015; Hu et al. 2016), and image captioning (Lu et al. 2018; Dai, Fidler, and Lin 2018). There have been two major lines of models for phrase grounding: (1) The two-stage phrase grounding models (Wang et al. 2018; Plummer et al. 2015, 2018; Chen et al. 2017a) first obtain candidate proposals from an explicit object detector and then perform matching according to their similarities to the query. (2) Single-stage phrase grounding models (Sadhu, Chen, and Nevatia 2019; Yang et al. 2019) directly generate dense candidate proposals on sliced image regions with various resolutions and perform matching with the query.

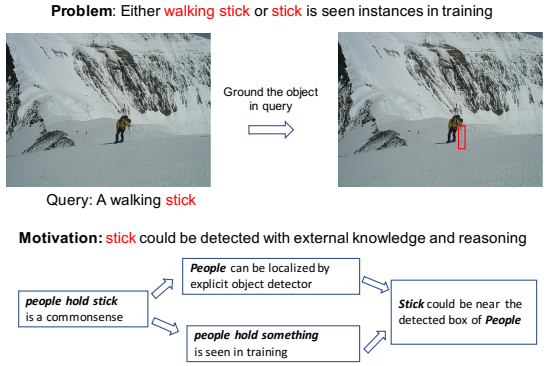

Figure 1: An example of zero-shot phrase grounding and motivation of our solution

There are two settings in the challenging zero-shot phrase grounding proposed by (Sadhu, Chen, and Nevatia 2019): (a) novel noun phrases, i.e., the noun phrases in the test set are not shown in the training set. (b) novel object categories, i.e., the visual examples of queried object categories in the test set are not shown in the training set, in which setting (b) is more strict than setting (a). To tackle the zero-shot problem, two-stage models can only ground a fixed set of object categories on which the explicit detector was trained, and the single-stage models lack object priors because they do not use explicit detectors. However, humans can effortlessly ground novel types of objects in natural language queries through reasoning based on knowledge (Minsky 2007). Inspired by this, we explore and propose a novel architecture by leveraging dense candidate proposals in single-stage models and relating the locations of the target proposal to objects detected by the explicit detector in a two-stage model, thus leveraging the best of both worlds. As illustrated in Figure 1, although neither *walking stick* nor *stick* is seen during training, we can ground them by resorting to commonsense knowledge "people-hold-stick" and learning to relate the target proposal to explicitly detected object box of "people".

To introduce external knowledge into our model, we propose to jointly learn from commonsense knowledge (Fellbaum 2012; Liu and Singh 2004) and scene graph knowledge (Krishna et al. 2017). We build multi-modal knowledge graphs based on text entities parsed from queries (**entity**

[*]Work done while interning at Samsung Research America
[†]Corresponding author

nodes), explicitly detected objects (**object nodes**) including their boxes (i.e., visual features and box coordinates), and class labels (i.e., semantic features, and retrieved relations (**relation nodes**) between entities and objects belonging to a pre-defined set of relation types). Once the multi-modal knowledge graph is built, we propose to perform graph reasoning by graph convolution operations to learn the context-aware representations for its nodes.

To enable spatial relation reasoning on the query, we need to relate the location of the target proposal to the explicitly detected object boxes. Between each detected object and the target proposal, we define eight types of spatial relations based on their intersection over union (IoU) values, relative distance d as well as relative angles θ: **Inside**, **Cover**, **Overlap**, **Top-Left**, **Top-Right**, **Bottom-Left**, **Bottom-Right** and **Irrelevant** as shown in Figure 3. The motivation for predicting a spatial relation between detected object boxes and target proposal is rendering a coarse estimate for the location of the noun phrases. We take as input the context-aware graph node features from the above multi-modal knowledge graphs to perform spatial relation reasoning.

Our approach is the first to incorporate external knowledge, graph reasoning, and spatial relation reasoning for zero-shot phrase grounding. We build on a baseline single-stage phrase grounding system (Sadhu, Chen, and Nevatia 2019), consisting of three major components: (1) Select the backbone encoder for images and candidate anchor box generator. (2) Perform multi-modal knowledge graphs (MMKG) reasoning to predict matching scores and regression parameters for each candidate proposal. (3) Perform spatial relation reasoning (SRR) with regard to detected objects to get localization scores for each candidate proposal. We take combined localization and matching scores as well as regression parameters as the final predictions. Extensive experiments were performed on zero-shot phrase grounding splits introduced by (Sadhu, Chen, and Nevatia 2019), which were developed on Visual Genome (Krishna et al. 2017) and Flickr30K Entities (Plummer et al. 2015; Young et al. 2014). Our models achieve significant improvement over the baseline single-stage phrase grounding model. Our main contributions are summarized as follows:

- We propose to construct multi-modal knowledge graphs based on external knowledge that connects queries and images and performs reasoning with graph convolution operations.
- A novel spatial relation reasoning component is developed to predict the spatial relation between target candidate proposal and detected boxes.
- Our proposed models show significant improvements over baselines on several zero-shot phrase grounding datasets. We provide detailed analyses on how these are achieved.

Related Work

Phrase Grounding There are two general frameworks for phrase grounding. The two-stage models (Plummer et al. 2015; Wang et al. 2018; Plummer et al. 2018; Chen et al. 2017a; Rohrbach et al. 2016; Yu et al. 2018, 2016; Mao et al. 2016) leverage an object detector such as FasterRCNN (Ren et al. 2015) and MaskR-CNN (He et al. 2017) in the first stage to obtain the bounding boxes and ROI-pooled features and then rank/classify the proposals in the second step. However, single-stage models (Sadhu, Chen, and Nevatia 2019; Yang et al. 2019; Zhao et al. 2018; Yeh et al. 2017; Yang, Li, and Yu 2020; Yang et al. 2020) instead use dense candidate proposals and directly fuse the text features from queries and visual representation from proposals to make the prediction. The dense candidate proposal features are usually from single-stage object detection such as SSD (Liu et al. 2016), Yolov3 (Redmon and Farhadi 2018), FPN (Lin et al. 2017a), RetinaNet (Lin et al. 2017b) and multiple other vision tasks (Yeh et al. 2017). The setting of zero-shot phrase grounding is first explored in (Sadhu, Chen, and Nevatia 2019), which proposes several zero-shot dataset splits subsampled from Flickr30K Entities (Plummer et al. 2015) and Visual Genome (Krishna et al. 2017).

External Knowledge Leveraging priors from external knowledge has been applied in both language and vision domains, e.g., language inference (Chen et al. 2017b), visual question answering (Singh et al. 2019; Li, Wang, and Zhu 2020), image classification (Marino, Salakhutdinov, and Gupta 2016), visual relation detection (Lu et al. 2016), object detection (Singh et al. 2018), commonsense reasoning (Ruan et al. 2019), etc. However, there has been very limited work on using reasoning with external knowledge for phrase grounding, especially in the challenging zero-shot setting where visual examples of test queries are not shown in training. We are the first to explore external knowledge, graph reasoning, and spatial relation reasoning for zero-shot phrase grounding. The work (Singh et al. 2018) on zero-shot detection is most similar to ours. However, their work directly encodes the external knowledge into image region proposals to perform classification where region proposals are already given as input and categories are predefined. Our approach aims to use external knowledge to localize the regions of phrases in sentences instead of classifying the specific proposed regions.

The Model

Given an image I and query q, the goal of phrase grounding is to localize the bounding box b_{gt} of the object referred by q in I. An overview of our phrase grounding framework is depicted in Figure 2, with the details of the components described in the following sections.

Image Encoder and Candidate Proposals

Following (Sadhu, Chen, and Nevatia 2019; Lin et al. 2017b), we apply nine candidate proposals of various scales and ratios on every sliced image region of different resolutions. To produce K feature maps $\{v_k\}_{k=1}^{K}$ at different resolutions, we use backbone image encoder ResNet-50 (He et al. 2016) with FPN (Lin et al. 2017a) as the default image encoder. Moreover, we perform l2-normalization along the channel dimension on their feature maps. Specifically, we denote visual feature $v_k[x, y]$ for a sliced image region with its center indexed by $[x, y]$ at the k^{th} feature map.

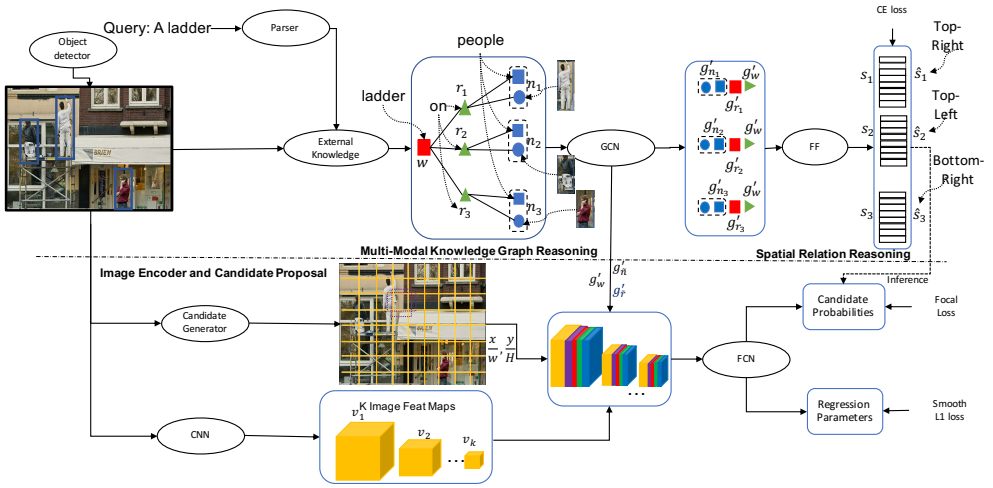

Figure 2: An overview of the proposed phrase grounding framework.

For one candidate proposal b_i, its final prediction is a five-dimensional vector; four dimensions are regression parameters z_i and one dimension is a confidence score p_i. Specifically, regression parameters z_i represent the shifts for the width, height, and center of b_i, while p_i is the combined score indicating the confidence on this shifted candidate proposal.

Multi-Modal Knowledge Graph (MMKG) Reasoning

A general challenge in building connections between unseen and seen object categories lies in obtaining a shared semantic representation space in the multi-modal setting.

Constructing MMKG We employ detectron2 (Wu et al. 2019) trained on the MSCOCO object detection classes (Lin et al. 2014) to recognize instances of objects and return their bounding boxes and object categories. As MSCOCO object classes may have direct overlap with queried nouns in validation or test set, we pick a subset of bounding boxes and their object classes as nodes of MMKG to avoid violating the zero-shot setting (b)[1]. We further encode these object boxes using our backbone image encoder and obtain Region of Interest (RoI) features, obtaining a set of visual features and instance category pairs $\{(v'_1, c_1), \cdots, (v'_M, c_M)\}$ where M is the number of detected objects and v', c are visual features and object classes, respectively. On the text side, we extract the lemmatized entity word w in the query.

Commonsense Knowledge We first consider building an undirected graph $\mathcal{G}_{cms}$ for a given image I and query q based on the commonsense knowledge from WordNet (Fellbaum 2012) and ConceptNet (Liu and Singh 2004), where the node set includes three types of nodes: **entity nodes**, **relation nodes**, and **object nodes**. Specifically, we use a lemmatized entity word w as the **entity node**, $n_i = (v'_i, c_i)$ as the i^{th} **object node** covering visual features and class labels, and r_i as a **relation node** that connects w and n_i. We add

[1]Both settings will not be violated since setting (b) is stricter than (a)

Wordnet	ConceptNet
Hypernymy	HasA
Hyponymy	InstanceOf, Entails IsA, MannerOf, DerivedFrom MadeOf, PartOf, TypeOf
Co-hyponyms	DistinctFrom
Synonymy	FormOf, SimilarTo, Synonym
N/A	AtLocation, LocatedNear, RelateTo

Table 1: Filtered Commonsense Relations

object nodes and **entity nodes** first and then **relation nodes** as well as edges based on knowledge triples; i.e., (w, r_i, n_i) will add a relation node r_i and assign two edges from node w to r_i and from r_i to n_i, respectively. Note that a relation node r_i could contain multiple relation types r_i^j and there would be no edges nor relation nodes between w and n_i, if there are no commonsense relations belonging to the set. We filter the massive commonsense knowledge relations to a pre-defined set $\{r\}$ critical to phrase grounding as shown in Table 1.

Scene Graph Knowledge To include more useful relation nodes between entity nodes and object nodes, we also explore semantic relations from textual scene graphs in Visual Genome (Krishna et al. 2017), e.g., "people-**on**-ladder" and "people-**carry**-papers". These semantic relations can provide additional prior for localization, which could be a good supplement to relation types from commonsense knowledge. Similarly, we construct a graph $\mathcal{G}_{sgg}$ based on scene graph knowledge, of which entity nodes and object nodes, as well as building process are the same with that of $\mathcal{G}_{cms}$. Note that we only preserve at most two semantic relations (based on frequency) in relation nodes between w and n_i from all scene graphs in Visual Genome.

To sum up, we build a multi-modal knowledge graph (MMKG) $\mathcal{G} = \mathcal{G}_{sgg} \cup \mathcal{G}_{cms}$ covering both visual and textual features based on both commonsense and scene graph knowledge.

MMKG Reasoning To perform reasoning on the con-

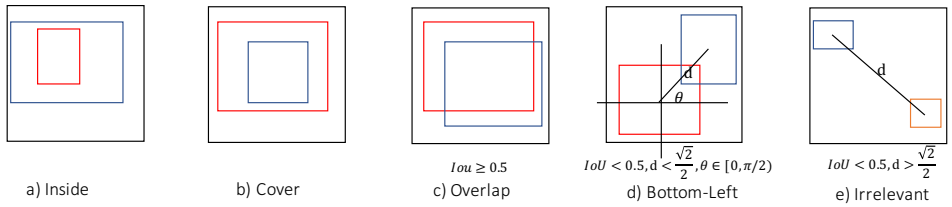

Figure 3: Spatial Relations between one object box (red) and target proposal (blue) on a scaled image .

structed MMKG, we use graph convolution operations on MMKG in the context of both modalities, images and text. Specifically, we initialize the features $g^0_{n_i}, g^0_{r_i}, g^0_w$ for object n_i, relation r_i and entity node w shown as follows:

$$g^0_{n_i} = f_n([v'_i; \mathbf{e}_{c_i}]) \tag{1}$$

$$g^0_{r_i} = \frac{1}{J}\sum_{j=1}^{J} \mathbf{e}^j_{r_i} \tag{2}$$

$$g^0_w = h \tag{3}$$

where f_n is feed-forward networks using the ReLU activation; $\mathbf{e}$ is the word embedding network; J is the number of relation types in r_i; h is normalized last hidden state vector with the query q encoded by Bi-LSTM.

We present the process of encoding $\mathcal{G}$ to produce a new set of context-aware representation. The representation of graph node r_i, n_i and w in the j^{th} convolution turn is computed as follows:

$$g^j_{r_i} = f_r([g^{j-1}_{n_i}; g^{j-1}_w; g^{j-1}_{r_i}]) \tag{4}$$

$$g^j_w = \frac{1}{N}\left[\sum_{r_i \in \mathcal{N}(w)} f_w([g^{j-1}_w; g^{j-1}_{r_i}])\right] \tag{5}$$

$$g^j_{n_i} = f_o([g^{j-1}_{n_i}, g^{j-1}_{r_i}]) \tag{6}$$

where f_r, f_o, f_w are feed-forward networks with the ReLU activation. $\mathcal{N}(w)$ denotes the adjacent nodes of w. N is the total number of adjacent nodes. We denote g'_{r_i}, g'_w, g'_{n_i} as representation after the graph convolution operations (2 convolution turns here).

Candidate Proposal Output To predict matching score $\tilde{p}_i$ and regression parameters z_i for candidate proposal b_i indexed at $[x, y]$ on feature map v_k, we use feed-forward network f_p with ReLU activation:

$$(\tilde{p}_i, z_i) = f_p([v_k[x, y], g'_{\tilde{r}}, g'_{\tilde{n}}, g'_w, x/W, y/H]) \tag{7}$$

where $g'_{\tilde{r}}$ and $g'_{\tilde{n}}$ are mean pooled features from relation and object nodes, respectively, and $x/W, y/H$ are scaled coordinates.

Spatial Relation Reasoning (SRR)

To make use of locations from detected objects, we explore the spatial relations between the ground truth boxes and the object boxes. The spatial relations represent their relative geometrical positions so that eight types of spatial relations are defined based on their Intersection over Union (IoU), relative distance d, and angle θ.

Specifically, given ground truth box b_{gt} and detected object n_i, the locations of them are denoted as (x_t, y_t) and (x_i, y_i), which are the normalized coordinates of the bounding box center, respectively. We can achieve the IoU between n_i and b_{gt}, relative angle θ and distance d. As shown in Figure 3, we first consider two types of special cases: **Cover** (n_i completely includes b_{gt}) and **Inside** (n_i is fully covered by b_{gt}). Apart from the two special types, if the IoU between b_{gt} and n_i is larger than 0.5, the geometrical relation is classified as **Overlap**. When the proportion of their relative distance d to image diagonal length is smaller than 0.5 and the IoU smaller than 0.5, we define the relation solely on quadrant of relative angle θ, namely **Top-Left**, **Top-Right**, **Bottom-Left** and **Bottom-Right**. Lastly, When the proportion is larger than 0.5 and IoU smaller than 0.5, the spatial relation between them is referred to as **Irrelevant**.

Following these rules, we could obtain ground truth spatial relation $\hat{s}_j$ between each detected object box n_j and the ground truth box b_{gt}. Meanwhile, we take concatenated features from graph nodes to reason about the spatial relations n_j between b_{gt} by predicting the probability distribution $s_j, s_j \in R^8$ over eight spatial relation categories by a feed-forward network f_s, and thus we use $\hat{s}_j$ as supervision for s_j in the SRR training.

$$s_j = f_s([g'_{n_j}; g'_w; g'_{r_j}]) \tag{8}$$

$$\hat{s}_j = \delta(b_{gt}, n_j) \tag{9}$$

where δ returns a one-hot spatial relation by the above rules.

SRR can be used to estimate the localization score p'_i of a given candidate proposal b_i with regard to a set of detected objects $\{n_j\}$ during inference:

$$p'_i = \frac{1}{M}\sum_j s_j^{\arg\max(\delta(b_i, n_j))} \tag{10}$$

where M is the number of detected objects and s^i_j will return the scalar indexed by i in vector s_j.

Training and Inference

During training, for all dense candidate proposals $B = \{b_i\}$, we obtain their corresponding matching scores set $\tilde{P} = \{\tilde{p}_i\}$ and regression parameters $Z = \{z_i\}$ from Equation 7 as well as a set of localization scores $P' = \{p'_i\}$ from Equation 10. Given the ground truth bounding box b_{gt}, we acquire a binary annotation set $T = \{t_i\}$ where $t_i = 1_{IoU(b_i, b_{gt}) \geq 0.5}$ is the indicator function denoting whether the IoU between one proposal b_i and b_{gt} exceeds 0.5. The losses can be writ-

	Train		Validation		Test	
	$\#i$	$\#q$	$\#i$	$\#q$	$\#i$	$\#q$
Flickr30K	30K	58K	1K	14K	1K	14K
Flickr-0	19K	11K	6K	9K	6K	9K
Flickr-1	19K	87K	6K	26K	6K	26K
VG-2UB	40K	251K	33K	83K	17K	23K
VG-2B	40K	251K	33K	83K	10K	12K
VG-3UB	40K	251K	33K	83K	41K	68K
VG-3B	40K	251K	33K	83K	23K	25K

Table 2: Dataset details, $\#i/\#q$ means image/query numbers

ten as below:

$$L_{cls} = \frac{1}{|T|}\sum_{i=1}^{|\tilde{P}|} L_F(\tilde{p}_i, t_i) \tag{11}$$

$$L_{reg} = \sum_{i=1}^{|Z|} t_i L_S(z_i, b_{gt}) \tag{12}$$

$$L_{loc} = \frac{1}{M}\sum_{i=1}^{M} L_C(s_i, \hat{s}_i) \tag{13}$$

where we use focal loss L_F ($\alpha = 0.25, \gamma = 2$) for the binary classification, smooth-L1 loss L_S for regression parameter predictions, and cross entropy loss L_C between predicted spatial relations s_i and ground truth $\hat{s}_i$. Therefore the overall loss can be written as $L = L_{cls} + \lambda_1 L_{reg} + \lambda_2 L_{loc}$. ($\lambda_1$ and λ_2 are set to be 1 here).

During inference, we derive the final confidence score $p_i = \tilde{p}_i + \beta p'_i$ and regression parameters z_i for a candidate anchor b_i, of which β is an adjustable hyper-parameter. We take the candidate proposal with the highest p as well as its regression parameters to get the predicted bounding box.

Experiments

Data Following (Sadhu, Chen, and Nevatia 2019), we use sub-sampled dataset split Flickr-0 and Flickr-1 from Flickr30K Entities (Plummer et al. 2015) with region-phrase correspondence annotated on the original Flickr30K (Young et al. 2014). We also use VG-2UB, VG-2B, VG-3UB, and VG-3B sub-sampled from Visual Genome (Krishna et al. 2017). The above six splits describe four slightly different cases of zero-shot phrase grounding settings.

Flickr-0 Flickr-0 follows zero-shot setting (a) that queried noun phrases in the test set are not included in training set; e.g., the noun phrase "blue sedan" would be regarded as zero-shot only if "sedan" are not included in training; Therefore noun phrases, such as "red minivan" and "old truck" are allowed in the training set although "sedan", "minivan" and "truck" belonging to the same object category, i.e., *vehicles*. Note that all noun phrases in Flickr30K Entities can be classified as 8 general categories, i.e., *people*, *clothing*, *bodyparts*, *animals*, *vehicles*, *instructs*, *scene*, and *other*.

Flickr-1 Flickr-1 follows zero-shot setting (b) The categories of nouns phrases in the test set are not included in the training set; e.g., the noun phrase "blue automobile" would be regarded as zero-shot only if visual examples of any vehicles such as "sedan" and "truck", which belong to the general category *vehicles*, are not included in the training.

VG-2UB, VG-2B VG-2UB and VG-2B follow zero-shot setting (b). Different from the Flickr dataset, Visual Genome defines more fine-grained object categories for noun phrases. UB and B mean unbalanced and balanced instance numbers in terms of object categories in testing, respectively.

VG-3UB, VG-3B VG-3UB and VG-3B follow zero-shot setting (b). Different from VG-2UB and VG-2B, these two splits include an image with at least one distracted object category; e.g., one test image would contain both a visual instance of unseen object category "bull" which is referenced by the query and a visual instance of seen object category "horse" to check if the model would tend to ground to the seen object category.

In addition to the zero-shot grounding dataset splits, we also conducted experiments on the standard split of Flikcr30K Entity.

More Experiment Set-up

Evaluation Metric We use the Intersection over Union (IoU) as in (Chen, Kovvuri, and Nevatia 2017). In the annotations of the above six zero-shot grounding dataset splits, there is exactly one ground truth box corresponding to a textual query. If the IoU of the predicted and the ground truth box is more than 0.5, we view it as the correct prediction. We calculate the number of correct predictions divided by the size of the test set to obtain accuracy.

Model Comparison We compare our models with the following baseline models: (1) QRG (Chen, Kovvuri, and Nevatia 2017) employs query-guided regression network by reinforcement learning for proposal generation and policy learning; we had its first-stage object detector pre-trained on Pascal-VOC (Everingham et al. 2010) and then fine-tuned on the training set of the specific dataset. (2) ZSG (Sadhu, Chen, and Nevatia 2019) combines the detector and text query to produce classification probabilities and regression parameters, which is the basic framework we build on. Note that since the method for YOLOG (Yang et al. 2019) is similar to ZSG except for the backbone encoder, we view it as the same baseline as ZSG.

Implementation Details To ensure a fair comparison, we follow the same setting as in (Sadhu, Chen, and Nevatia 2019) in terms of text encoders, backbone image encoders, and the optimization method. Specifically, we use the Glove embedding (Pennington, Socher, and Manning 2014) and Bi-LSTM (hidden dimension 256) for query features, and SSD (Liu et al. 2016) with VGG16 network or RetinaNet (Lin et al. 2017b) with Resnet-50 (He et al. 2016) network initialized with features pre-trained on ImageNet (Deng et al. 2009) for backbone image encoder. The hyper-parameters $\lambda_1, \lambda_2, \beta$ are set to be 1, 1 and 0.5, respectively. The feed-forward networks f_n, f_r, f_w, f_o in MMKG and the SRR module are two-layer feed-forward networks with the output dimension being 256. We perform graph convolution operations twice to get contextual-

ized representation. Same as in (Sadhu, Chen, and Nevatia 2019), we start training by resizing the image to 300 * 300 for 10 epochs, and then we fine-tune the network with images being resized to 600 * 600 for 20 epochs using Adam (Kingma and Ba 2014) with a learning rate of $1e^{-4}$. Image augmentation methods such as flipping are not used because they may change the relative spatial relations here. For Flickr and Visual Genome sub-sampled split, we pick bounding boxes from the explicit object detector belonging to the Flickr30K or Visual Genome categories. In the Flickr sub-sampled splits, we further replace fine-grained MSCOCO object classes such as "cat" and "dog" with general Flickr30K category "animal" when returning detected bounding box labels.

Quantitative Analysis

Overall Performance Table 3 and 4 show the results on the zero-shot learning splits and Flickr30K Entity standard split. The tables show that our proposed method outperforms the baselines on the standard metrics. Our model reaches the accuracy of 45.99%, 33.12%, 15.12%, 14.49%, 15.93%, and 15.28% on the six zero-shot learning splits, achieving an improvement of 2 $\sim$ 3%. Similar to the baseline methods, the accuracy on Flickr-0,1 is much higher than on the VG-split, which is partially due to the higher variance of categories of the referred objects in Visual Genome. Our method is orthogonal to backbones, and our results using VGG16 as backbone are also much higher than the baseline. On the standard Flickr30K Entity split, our model reaches 65.02% and 62.56%, yielding an improvement of 1.7% and 2.4% over the baseline methods using different backbones.

Ablation Analysis Table 3 shows ablation analyses on different component compositions. The results show that both the multi-modal knowledge graph reasoning (denoted as G) and spatial relation reasoning (denoted as R) help improve the performance of the phrase grounding; e.g., "Base + G" increases "Base" from 31.23% to 32.33% and "Base + G + R" further improves the performance from 32.33% to 33.12% on Flickr-1. Interestingly, while the scene graph knowledge added model "Base + G_{sgg}" improves the baseline by a relative small margin (43.02% to 43.18% on Flickr-0) compared to the commonsense knowledge graph "Base + G_{cms}" (43.02% to 44.10%), it could work as a useful supplement to G_{cms} to further boost performance (44.10% to 44.41%). Hence we use the union of both knowledge $G = G_{cms} \cup G_{sgg}$ in the final model.

Reasoning Module Analyses We analyze the performance of the reasoning per spatial relation and list the effect of hyper-parameter β on the final prediction. Figure 4 illustrates the precision and recall of every spatial relation category, showing that the module achieves relatively higher precision than recall, particularly in terms of predicting "Irrelevant", "Cover" and "Overlap" and thus providing strong and accurate cues to localize the objects referred by a query. One more interesting point is that the reasoning module tends to reach a higher recall on direction based spatial relations, e.g., "Top-Left" and "Top-Right", because these two relations belong to the most common categories among the eight.

In Figure 5, our performance reaches its peak with β being set to be around 0.5 on Flickr-0,1 validation set. Moreover, our results will be strongly boosted when we use ground truth reasoning results instead of the model predictions, reaching 48.64% with β set to be 1. It demonstrates that spatial relation reasoning is an effective auxiliary task for zero-shot phrase grounding.

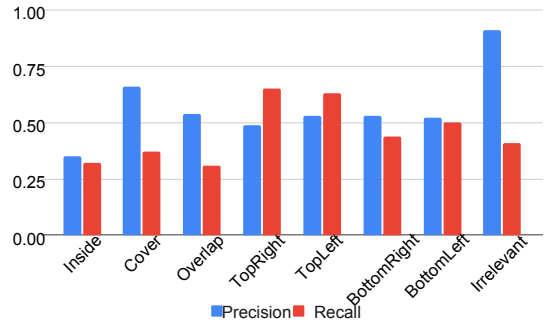

Figure 4: Result of spatial relations on Flickr-1 test set.

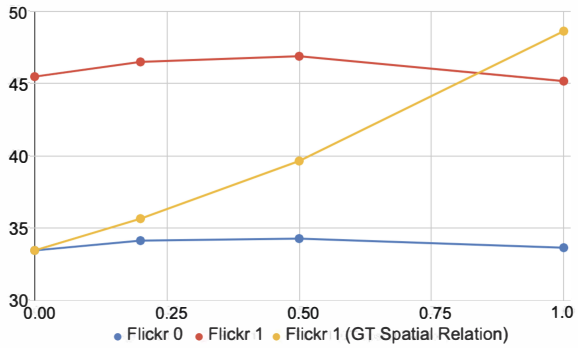

Figure 5: Effect of β on validation set of two split.

Knowledge Module Analysis We analyze how nodes in multi-modal graphs affect the performance. On Flickr-0,1, we compare results by removing features from object nodes and relation nodes. As shown in Table 5, relation nodes contribute significantly to the prediction (Flickr-0 results dropped from 45.99% to 43.48% and Flickr-1 dropped from 33.12% to 31.52%) and visual features of detected object boxes also affect the performance dramatically (Flickr-0 decreased from 45.99% to 44.11% and Flickr-1 decreased from 33.12% to 32.49%).

Qualitative Analysis

Case Study Figure 6 shows six zero-shot phrase grounding examples benefiting from the detected bounding boxes of "people" from explicit object detector and external knowledge extracted upon the query and "people". While baseline methods either ground novel phrases to partial of the bounding boxes of irrelevant but seen objects in training (the bottom three examples) or simply to image background (the top three examples), our methods show better performance in grounding the unseen objects in text queries. All the six examples can leverage cues from knowledge triplets and the detected bounding box of "people", and hence unseen noun phrases would be localized in a more precise way.

	BackBone	Flickr-0	Flickr-1	VG-2B	VG-2UB	VG-3B	VG-3UB
QRG (Chen, Kovvuri, and Nevatia 2017)	VGG16	35.62	24.42	7.64	7.15	8.35	7.52
ZSG (Sadhu, Chen, and Nevatia 2019)	VGG16	39.32	29.35	11.02	10.55	11.42	10.97
Base + G + R (This work)	VGG16	**42.88**	**31.41**	**13.54**	**12.99**	**14.12**	**13.40**
ZSG (Sadhu, Chen, and Nevatia 2019)	ResNet50	43.02	31.23	12.90	12.37	13.77	12.82
Base + G_{cms}	ResNet50	44.10	32.33	14.19	13.68	15.03	14.18
Base + G_{sgg}	ResNet50	43.18	31.47	13.90	13.39	14.79	13.61
Base + G	ResNet50	44.41	32.42	14.43	13.81	15.31	14.49
Base + R	ResNet50	43.48	31.52	13.42	12.76	14.16	13.37
Base + G_{cms} + R	ResNet50	45.72	33.31	14.96	14.30	15.61	14.76
Base + G_{sgg} + R	ResNet50	44.29	32.33	14.31	13.95	15.23	14.35
Base + G + R (This work)	ResNet50	**45.99**	**33.12**	**15.12**	**14.49**	**15.93**	**15.28**

Table 3: Accuracy (%) on zero-shot dataset splits

	BackBone	Flickr30K
SCRC (Hu et al. 2016)	VGG16	27.80
GroundR (Rohrbach et al. 2016)	VGG16	48.38
GroundR (Rohrbach et al. 2016)	VGG16	42.43
MCB (Fukui et al. 2016)	VGG16	48.70
CITE* (Plummer et al. 2018)	VGG16	61.89
QRG* (Chen, Kovvuri, and Nevatia 2017)	VGG16	60.10
ZSG (Sadhu, Chen, and Nevatia 2019)	VGG16	60.12
This work	VGG16	**62.56**
ZSG (Sadhu, Chen, and Nevatia 2019)	ResNet50	63.39
This work	ResNet50	**65.02**

Table 4: Accuracy (%) on the Flickr30K Entity dataset. Models with "*" have their first-stage object detector fine-tuned on Flickr30k Entity objects

	Flickr-0	Flickr-1
This work	45.99	33.12
-w/o visual features in object nodes	44.11	32.49
-w/o textual featues in object nodes	45.51	32.91
-w/o relation node	43.48	31.52

Table 5: Effect of different types of node features and limited categories of explicit object detector.

Failure Cases We also investigate several scenarios where our method does not have any improvement compared to baselines in the zero-shot phrase grounding: (1) Lack of specific knowledge. In Figure 7 (a) and (b), we need to model more specific knowledge about "conductor" and "judge" so that we can localize them correctly from other "people" instances. (2) Multiple objects. Our model does not differentiate singular or plural forms of text phrases when building MMKG, thus having little improvement on these cases, as shown in Figure 7 (c) and (d). (3) Vague or general noun phrases. Our model tends to ground a large area of the image, same with the baseline in these cases, as shown in Figure 7 (e) and (f).

Conclusions

This paper explores better solutions for the fundamental problem of zero-shot phrase grounding by building multi-modal knowledge graphs based on external knowledge and performing two types of reasoning based on the acquired

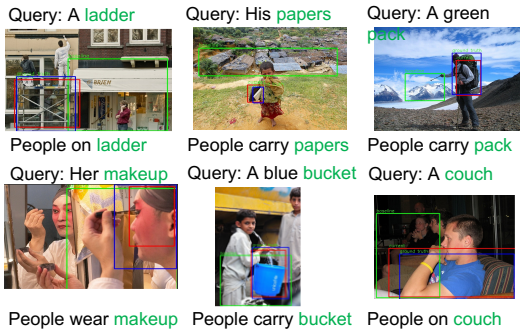

Figure 6: Zero-shot Ground Cases from Flickr-0,1. Green, blue and red boxes denote localization by baseline, ground truth and our models, respectively

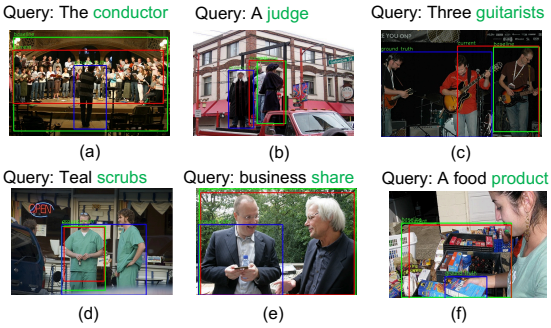

Figure 7: Zero-shot Grounding Failure Cases.

graphs. Specifically, we connect seen, and unseen categories of objects referenced by text queries using multi-modal knowledge graphs (MMKG) built with commonsense and scene graph knowledge. We perform both MMKG reasoning and spatial relation reasoning to localize noun phrases referenced by queries. Our experiments on different zero-shot grounding datasets sub-sampled from Flickr30K Entities and Visual Genome show that the proposed model significantly outperforms baselines.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual

question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Chen, K.; Kovvuri, R.; Gao, J.; and Nevatia, R. 2017a. MSRC: Multimodal spatial regression with semantic context for phrase grounding. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 23–31.

Chen, K.; Kovvuri, R.; and Nevatia, R. 2017. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, 824–832.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Inkpen, D.; and Wei, S. 2017b. Neural natural language inference models enhanced with external knowledge. *arXiv preprint arXiv:1711.04289*.

Dai, B.; Fidler, S.; and Lin, D. 2018. A neural compositional paradigm for image captioning. In *Advances in Neural Information Processing Systems*, 658–668.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Fellbaum, C. 2012. WordNet. *The encyclopedia of applied linguistics*.

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*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6904–6913.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, R.; Xu, H.; Rohrbach, M.; Feng, J.; Saenko, K.; and Darrell, T. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4555–4564.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3668–3678.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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. *International journal of computer vision*, 123(1): 32–73.

Li, G.; Wang, X.; and Zhu, W. 2020. Boosting Visual Question Answering with Context-aware Knowledge Aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1227–1235.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

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 *European conference on computer vision*, 740–755. Springer.

Liu, H.; and Singh, P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4): 211–226.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *European conference on computer vision*, 852–869. Springer.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7219–7228.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Marino, K.; Salakhutdinov, R.; and Gupta, A. 2016. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*.

Minsky, M. 2007. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Plummer, B. A.; Kordas, P.; Hadi Kiapour, M.; Zheng, S.; Piramuthu, R.; and Lazebnik, S. 2018. Conditional image-text embedding networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 249–264.

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 *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 91–99.

Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, 817–834. Springer.

Ruan, Y.-P.; Zhu, X.; Ling, Z.-H.; Shi, Z.; Liu, Q.; and Wei, S. 2019. Exploring unsupervised pretraining and sentence structure modelling for winograd schema challenge. *arXiv preprint arXiv:1904.09705*.

Sadhu, A.; Chen, K.; and Nevatia, R. 2019. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE International Conference on Computer Vision*, 4694–4703.

Singh, A. K.; Mishra, A.; Shekhar, S.; and Chakraborty, A. 2019. From strings to things: Knowledge-enabled VQA model that can read and reason. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4602–4612.

Singh, K. K.; Divvala, S.; Farhadi, A.; and Lee, Y. J. 2018. Dock: Detecting objects by transferring common-sense knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 492–508.

Wang, L.; Li, Y.; Huang, J.; and Lazebnik, S. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 394–407.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Yang, S.; Li, G.; and Yu, Y. 2020. Propagating Over Phrase Relations for One-Stage Visual Grounding. In *European Conference on Computer Vision*, 589–605. Springer.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving one-stage visual grounding by recursive sub-query construction. *arXiv preprint arXiv:2008.01059*.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, 4683–4693.

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 *Advances in Neural Information Processing Systems*, 1912–1922.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1307–1315.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, 69–85. Springer.

Zhao, F.; Li, J.; Zhao, J.; and Feng, J. 2018. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5696–5705.