Grounding Referring Expressions in Images by Variational Context

Hanwang Zhang[1*] Yulei Niu[2†] Shih-Fu Chang[3]
[1]Nanyang Technological University, [2]Renmin University of China, [3]Columbia University,
hanwangzhang@ntu.edu.sg; niu@ruc.edu.cn; shih.fu.chang@columbia.edu

Abstract

We focus on grounding (i.e., localizing or linking) referring expressions in images, e.g., "largest elephant standing behind baby elephant". This is a general yet challenging vision-language task since it does not only require the localization of objects, but also the multimodal comprehension of context — visual attributes (e.g., "largest", "baby") and relationships (e.g., "behind") that help to distinguish the referent from other objects, especially those of the same category. Due to the exponential complexity involved in modeling the context associated with multiple image regions, existing work oversimplifies this task to pairwise region modeling by multiple instance learning. In this paper, we propose a variational Bayesian method, called Variational Context, to solve the problem of complex context modeling in referring expression grounding. Our model exploits the reciprocal relation between the referent and context, i.e., either of them influences estimation of the posterior distribution of the other, and thereby the search space of context can be greatly reduced. We also extend the model to unsupervised setting where no annotation for the referent is available. Extensive experiments on various benchmarks show consistent improvement over state-of-the-art methods in both supervised and unsupervised settings. The code is available at https://github.com/yuleiniu/vc/.

1. Introduction

Grounding natural language in visual data is a hallmark of AI, since it establishes a communication channel between humans, machines, and the physical world, underpinning a variety of multimodal AI tasks such as robotic navigation [38], visual Q&A [1, 16, 49], and visual chatbot [6]. Thanks to the rapid development in deep learning-based CV and NLP, we have witnessed promising results not only in grounding nouns (*e.g.*, object detection [30]),

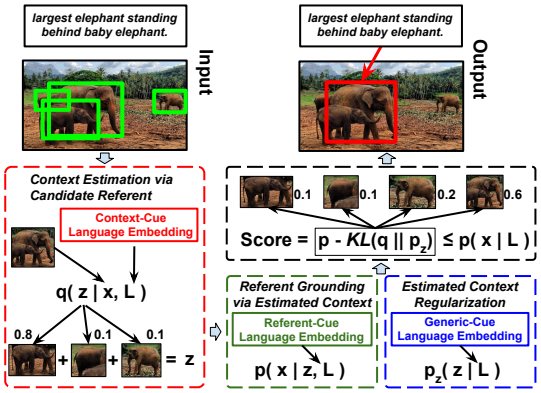

Figure 1. The proposed Variational Context model. Given an input referring expression and an image with region proposals, we localize the referent as output. We develop a grounding score function, with the variational lower-bound composed by three cue-specific multimodal modules, indicated by the description in the dashed color boxes.

but also short phrases (*e.g.*, noun phrases [28] and relations [47, 36]). However, the more general task: grounding referring expressions [25], is still far from resolved due to the challenges in understanding of both language and scene compositions [10]. As illustrated in Figure 1, given an input referring expression "largest elephant standing behind baby elephant" and an image with region proposals, a model that can only localize "elephant" is not satisfactory as there are multiple elephants. Therefore, the key for referring expression grounding is to comprehend and model the context. Here, we refer to *context* as the visual objects (*e.g.*, "elephant"), attributes (*e.g.*, "largest" and "baby"), and relationships (*e.g.*, "behind") mentioned in the expression that help to distinguish the referent from other objects.

One straightforward way of modeling the relations between the referent and context is to: 1) use external syntactic parsers to parse the expression into entities, modifiers, and relations [34], and then 2) apply visual relation detectors to localize them [47]. However, this two-stage approach is not practical due to the limited generalization ability of the detectors applied in the highly unrestricted language and scene compositions. To this end, re-

*Hanwang was a research scientist at Columbia University.
†Yulei was a visiting student at Columbia University.

cent approaches use multimodal embedding networks that jointly comprehend language and model the visual relations [26, 11]. Due to the prohibitively high cost of annotating both referent and context of referring expressions in images, multiple instance learning (MIL) [7] is usually adopted in them to handle the weak supervision of the unannotated context objects, by maximizing the joint likelihood of every region pair. However, for a referent, the MIL framework essentially oversimplifies the number of context configurations of N regions from $\mathcal{O}(2^N)$ to $\mathcal{O}(N)$. For example, to localize the "elephant" in Figure 1, we may need to consider the other three elephants all together as a multinomial subset for modeling the context such as "largest", "behind" and "baby elephant".

In this paper, we propose a novel model called *Variational Context* for grounding referring expressions in images. Compared to the previous MIL-based approaches [26, 11], our model approximates the combinatorial context configurations with weak supervision using a variational Bayesian framework [15]. Intuitively, it exploits the reciprocity between referent and context, given either of which can help to localize the other. As shown in Figure 1, for each region x, we first estimate a coarse context z, which will help to refine the true localizations of the referent. This reciprocity is formulated into the variational lower-bound of the grounding likelihood $p(x|L)$, where L is the text expression and the context is considered as a hidden variable z (cf. Section 3). Specifically, the model consists of three multimodal modules: context posterior $q(z|x, L)$, referent posterior $p(x|z, L)$, and context prior $p_z(z|L)$, each of which performs a grounding task (cf. Section 4.3) that aligns image regions with a cue-specific language feature; each cue dynamically encodes different subsets of words in the expression L that help the corresponding localization (cf. Section 4.2).

Thanks to the reciprocity between referent and context, our model can not only be used in the conventional supervised setting, where there is annotation for referent , but also in the challenging unsupervised setting, where there is no instance-level annotation (*e.g.*, bounding boxes) of both referent and context. We perform extensive experiments on four benchmark referring expression datasets: RefCLEF [14], RefCOCO [45], RefCOCO+ [45], and RefCOCOg [25]. Our model consistently outperforms previous methods in both supervised and unsupervised settings. We also qualitatively show that our model can ground the context in the expression to the corresponding image regions (cf. Section 5).

2. Related Work

Grounding Referring Expressions. Grounding referring expression is also known as referring expression comprehension, whose inverse task is called referring expression generation [25]. Different from grounding phrases [29, 28] and descriptive sentences [12, 32], the key for grounding referring expression is to use the context (or pragmatics in linguistics [37]) to distinguish the referent from other objects, usually of the same category [10]. However, most previous works resort to use holistic context such as the entire image [25, 12, 32] or visual feature difference between regions [45, 46]. Our model is similar to the works on explicitly modeling the referent and context region pairs [11, 26], however, due to the lack of context annotation, they reduce the grounding task into a multiple instance learning framework [7]. As we will discuss later, this framework is not a proper approximation to the original task. There are also studies on visual relation detection that detect objects and their relationships [21, 5, 47, 17, 48]. However, they are limited to a fixed-vocabulary set of relation triplets and hence are difficult to be applied in natural language grounding. Our cue-specific language feature is similar to the language modular network [11] that learns to decompose a sentence into referent/context-related words, which are different from other approaches that treat the expression as a whole [25, 23, 46, 19].

Variational Bayesian Model vs. Multiple Instance Learning. Our proposed variational context model is in a similar vein of the deep neural network based variational autoencoder (VAE) [15], which uses neural networks to approximate the posterior distribution of the hidden value $q(z|x)$, *i.e.*, encoder, and the conditional distribution of the observation $p(x|z)$, *i.e.*, decoder. VAE shows efficient and effective end-to-end optimization for the *intractable* log-sum likelihood $\log \sum_z p(x, z)$ that is widely used in generative processes such as image synthesis [44] and video frame prediction [43]. Considering the unannotated context as the hidden variable z, the referring expression grounding task can also be formulated into the above log-sum marginalization (cf. Eq. (2)). The MIL framework [7] is essentially a sum-log approximation of the log-sum, *i.e.*, $\sum_z \log p(x, z)$. To see this, the max-pooling function $\log \max_z p(x, z)$ used in [11] can be viewed as the sum-log $\sum_z \log p(x|z)p(z)$, where $p(z) = 1$ if z is the correct context and 0 otherwise, indicating there is only one positive instance; maximizing the noisy-or function $\log(1 - \prod_z (1 - p(x, z)))$ used in [26] is equivalent to maximize $\sum_z \log p(x, z)$, assuming there is at least one positive instance. However, due to the numerical property of the log function, this sum-log approximation will unnecessarily force every (x, z) pair to explain the data [8]. Instead, we use the variational Bayesian upper-bound to obtain a better sum-log approximation. Note that visual attention models [2, 42] simplify the variational lower bound by assuming $p(z) = q(z|x)$; however, we explicitly use the KL divergence $KL(q(z|x)||p(z))$ in the lower bound to regularize the approximate posterior $q(z|x)$ not being too far from the prior $p(z)$.

3. Variational Context

In this section, we derive the variational Bayesian formulation of the proposed variational context model and the objective function for training and test.

3.1. Problem Formulation

The task of grounding a referring expression L in an image I, represented by a set of regions $x \in \mathcal{X}$, can be viewed as a region retrieval task with the natural language query L. Formally, we maximize the log-likelihood of the conditional distribution to localize the referent region $x^* \in \mathcal{X}$:

$$x^* = \arg\max_{x \in \mathcal{X}} \log p(x|L), \tag{1}$$

where we omit the image I in $p(x|I, L)$.

As there is usually no annotation for the context, we consider it as a hidden variable z. Therefore, Eq. (1) can be rewritten as the following maximization of the log-likelihood of the conditional marginal distribution:

$$x^* = \arg\max_{x \in \mathcal{X}} \log \sum_z p(x, z|L). \tag{2}$$

Note that z is NOT necessary to be one region as assumed in recent MIL approaches [11, 26], *i.e.*, $z \in \mathcal{X}$. For example, the contextual objects "surrounding elephants" in "a bigger elephant than the surrounding elephants" should be composed by a multinomial subset of $\mathcal{X}$, resulting in an extremely large sample space that requires $\mathcal{O}(2^{|\mathcal{X}|})$ search complexity. Therefore, the marginalization in Eq (2) is intractable in general.

To this end, we use the variational lower-bound [15] to approximate the marginal distribution in Eq. (2) as:

$$\log p(x|L) = \log \sum_z p(x, z|L) \geq \mathcal{Q}(x, L) = \underbrace{\mathbb{E}_{z \sim q_\phi(z|x,L)} \log p_\theta(x|z, L)}_{\text{Localization}} - \underbrace{KL\left(q_\phi(z|x, L)||p_\omega(z|L)\right)}_{\text{Regularization}}, \tag{3}$$

where $KL(\cdot)$ is the Kullback-Leibler divergence, ϕ, θ, and ω are independent parameter sets for the respective distributions. As shown in Figure 1, the lower bound $\mathcal{Q}(x, L)$ offers a new perspective for exploiting the reciprocal nature of referent and context in referring expression grounding:

Localization. This term calculates the localization score for x given an estimated context z, using the referent-cue of L parameterized by θ. In particular, we design a new posterior $q_\phi(z|x, L)$ that approximates the true context prior $p(z|x, L)$, which models the context z using the context-cue of L parameterized by ϕ. In the view of variational auto-encoder [15, 35], this term works in an encoding-decoding fashion: q_ϕ is the encoder from x to z, and p_θ is the decoder from z to x.

Regularization. As KL is non-negative, maximizing $\mathcal{Q}(x, L)$ would encourage that the posterior q_ϕ is similar to the prior p_ω, *i.e.*, the estimated context z sampled from $q_\phi(z|x, L)$ should not be too far from the referring expression, which is modeled by $p_\omega(z|L)$ with the generic-cue of L parameterized by ω. This term is necessary as the estimated z could be overfitted to region features that are inconsistent with the visual context described in the expression.

3.2. Training and Test

Deterministic Context. The lower-bound $\mathcal{Q}(x, L)$ transforms the intractable log-sum in Eq. (2) into the efficient sum-log in Eq. (3), which can be optimized by using Monte Carlo unbiased gradient estimator such as REINFORCE [40]. However, due to that ϕ is dependent on the sampling of z over $\mathcal{O}(2^{|\mathcal{X}|})$ configurations, its gradient variance is large. To this end, we implement $q_\phi(z|x, L)$ as a differentiable but biased encoder:

$$z = f(x, L) = \sum_{x' \in \mathcal{X}} x' \cdot q_\phi(x'|x, L), \tag{4}$$

where we slightly abuse q_ϕ as a score function such that $\sum_{x'} q_\phi(x'|x, L) = 1$. Note that this deterministic context can be viewed as applying the "re-parameterization" trick as in Variational Auto-Encoder [15]: rewriting $z \sim q_\phi(z|x, L)$ to $z = f(x, L; \epsilon), \epsilon \sim p(\epsilon)$, where the stochasticity of the auxiliary random variable ϵ comes from training samples $x \in \mathcal{X}(\epsilon)$. A clear example is Adversarial Autoencoder [24] which shows that such stochasticity achieves similar test-likelihood compared to other distributions such as Gaussian.

Objective Function. Applying Eq. (4) to Eq. (3), we can rewrite $Q(x, L)$ into a function of only one sample estimation, which is a common practice in SGD:

$$\mathcal{Q}(x,L) = \log p_\theta(x|z,L) - \log q_\phi(z|x,L) + \log p_\omega(z|L). \tag{5}$$

In supervised setting where the ground truth of the referent is known, to distinguish the referent from other objects, we need to train a model that outputs a high $p(x|L)$ (*i.e.*, $\mathcal{Q}(x, L)$), while maintaining a low $p(x'|L)$ (*i.e.*, $\mathcal{Q}(x', L)$), whenever $x' \neq x$. Therefore, we use the so-called Maximum Mutual Information loss as in [25] $-\log\{\mathcal{Q}(x, L)/\sum_{x'} \mathcal{Q}(x', L)\}$, where we do not need to explicitly model the distributions with normalizations; we use the following score function:

$$\mathcal{Q}(x, L) \propto \mathcal{S}(x, L) = s_\theta(x, L) - s_\phi(x, L) + s_\omega(x, L), \tag{6}$$

where z is omitted as it is a function of x in Eq. (4). s_θ, s_ϕ, and s_ω are the score functions (*e.g.*, $p_\theta \propto s_\theta$) for p_θ, q_ϕ, and p_ω, respectively. These functions will be detailed in

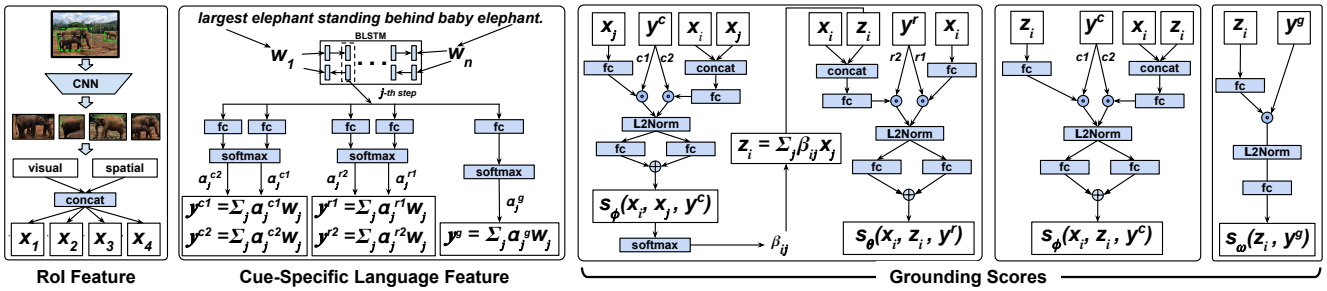

Figure 2. The architecture of the proposed Variational Context model. It consists of a region feature extraction module (Section 4.1, and a language feature extraction module (Section 4.2), and three grounding modules (Section 4.3). It can be trained in an end-to-end fashion with the input of a set of image regions and a referring expression, using the supervised loss (Eq. (7)) or the unsupervised loss (Eq. (8)). fc: fully-connected layer. concat: vector concatenation. L2Norm: L2 normalization layer. $\odot$: element-wise vector multiplication. $\oplus$: add.

Section 4.3. In this way, maximizing Eq. (5) is equivalent to minimizing the following softmax loss:

$$\mathcal{L}_s = -\log \text{softmax}\, \mathcal{S}(x_{gt}, L), \tag{7}$$

where the softmax is over $x \in \mathcal{X}$ and x_{gt} is the ground truth referent region.

Note that the reciprocity between referent and context can be extended to unsupervised learning, where neither of the referent and context has annotation. In this setting, we adopt the *image-level* max-pooled MIL loss functions for unsupervised referring expression grounding:

$$\mathcal{L}_u = -\max_{x \in \mathcal{X}} \log \text{softmax}\, \mathcal{S}(x, L), \tag{8}$$

where the softmax is over $x \in \mathcal{X}$. Note that the max-pooled MIL function is reasonable since there is only one ground truth referent given an expression and image training pair.

At test stage, in both supervised and unsupervised settings, we predict the referent region x^* by selecting the region $x \in \mathcal{X}$ with the highest score:

$$x^* = \arg\max_{x \in \mathcal{X}} \mathcal{S}(x, L), \tag{9}$$

4. Model Architecture

The overall architecture of the proposed variational context model is illustrated in Figure 2. Thanks to the deterministic context in Eq. (4), the five modules in our model can be integrated into an end-to-end differentiable fashion. Next, we will detail the implementation of each module.

4.1. RoI Features

Given an image with a set of Region of Interests (RoIs) $\mathcal{X}$, obtained by any off-the-shelf proposal generator [50] or object detectors [20], this module extracts the feature vector $\mathbf{x}_i$ for every RoI. In particular, $\mathbf{x}_i$ is the concatenation of visual feature $\mathbf{v}_i$ and spatial feature $\mathbf{p}_i$. For $\mathbf{v}_i$, we can use the output of a pre-trained convolutional network (cf. Section 5). If the object category of each RoI is available, we can further utilize the comparison between the referent and other objects to capture the visual difference such as "the largest/baby elephant". Specifically, we append the visual difference feature [45] $\delta\mathbf{v}_i = \frac{1}{n}\sum_{j \neq i} \frac{\mathbf{v}_i - \mathbf{v}_j}{||\mathbf{v}_i - \mathbf{v}_j||}$ to the original $\mathbf{v}_i$ visual feature, where n is the number of objects chosen for comparison (*e.g.*, the number of RoI in the same object category). For spatial feature, we use the 5-d spatial attributes $\mathbf{p}_i = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}]$, where x and y are the coordinates the top left (tl) and bottom right (br) RoI of the size $w \times h$, and the image is of the size $W \times H$.

4.2. Cue-Specific Language Features

The cue-specific language feature representation for a referring expression is inspired by the attention weighted sum of word vectors [11, 22, 3], where the weights are parameterized by context-cue ϕ, referent-cue θ, and generic-cue ω. The context-cue language feature $\mathbf{y}^c = [\mathbf{y}^{c1}, \mathbf{y}^{c2}]$ is a concatenation of $\mathbf{y}^{c1}$ for language-vision association between *single* RoI and the expression, and $\mathbf{y}^{c2}$ for the association between *pairwise* RoIs; the referent-cue language feature $\mathbf{y}^r$ can be represented in a similar way to $\mathbf{y}^c$; the generic-cue language feature $\mathbf{y}^g$ is only for single RoI association as it is an independent prior. The weights of each cue are calculated from the hidden state vectors of a 2-layer bidirectional LSTM (BLSTM) [33], scanning through the expression. The hidden states encode forward and backward compositional semantic meanings of the sentences, beneficial for selecting words that are useful for single and pairwise associations. Specifically, suppose $\mathbf{h}_j$ as the 4,000-d concatenation of forward and backward hidden vectors of the j-th word, without loss of generality, the word attention weight α_j and the language feature $\mathbf{y}$ for single/pairwise association of any cue can be calculated as:

$$\mathbf{m}_j = \text{fc}(\mathbf{h}_j), \alpha_j = \text{softmax}_j(\mathbf{m}_j), \mathbf{y} = \sum\nolimits_j \alpha_j \mathbf{w}_j, \tag{10}$$

where $\mathbf{w}_j$ is a 300-d vector. Note that the BLSTM module can be jointly trained with the entire model.

Figure 3 shows that the cue-specific language features dynamically weight words in different expressions. We can have two interesting observations. First, c1 is almost uni-

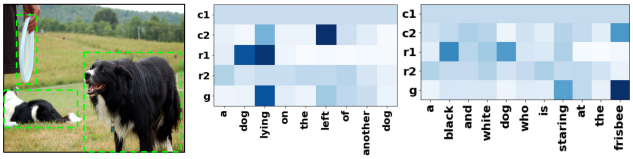

Figure 3. Two qualitative examples of the cue-specific language feature word weights. Darker color indicates higher weights. c/r+1/2: context/referent-cue + single/pairwise.

form while c2 is highly skewed; although r2 is more skewed than c1, it is still less skewed than r1. This is reasonable since: 1) without ground-truth, individual score (c1) does not help much for context estimation from scratch; context is more easily found by the pairwise score (c2) induced by relationships or other objects (*e.g.*, "left" or "frisbee"); 2) in referent grounding with ground truth, individual score (r1) is sufficient (*e.g.*, "dog lying" and "black white dog") and pairwise score (r2) is helpful; 3) g is adaptive to the number of object categories in the expression, *i.e.*, if the context object is of the same category as the referent, g weighs descriptive or relationship words higher (*e.g.*, "lying, standing, left"), and nouns higher (*e.g.*, "frisbee"), otherwise; moreover, it demonstrates that the deterministic guess of z in Eq. (4) is meaningful.

4.3. Score Functions

For any image and expression pair, given the RoI feature $\mathbf{x}_i$, and the cue-specific language feature $\mathbf{y}^c$, $\mathbf{y}^r$, and $\mathbf{y}^g$, we implement the final grounding score in Eq. (6) as:

$$\begin{aligned} \mathbf{z}_i = \sum\nolimits_j & \text{softmax}_j \left(s_\phi(\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}^c)\right) \mathbf{x}_j, \\ & s_\theta(x, L) \leftarrow s_\theta(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}^r), \\ & s_\phi(x, L) \leftarrow s_\phi(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}^c), \\ & s_\omega(x, L) \leftarrow s_\omega(\mathbf{z}_i, \mathbf{y}^g), \end{aligned} \tag{11}$$

where the right-hand side functions are defined as below.

Context Estimation Score: $s_\phi(\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}^c)$. It is a score function for modeling the context posterior $q_\phi(z|x, L)$, *i.e.*, given an RoI $\mathbf{x}_i$ as the candidate referent, we calculate the likelihood of any RoI $\mathbf{x}_j$ to be the context. We can also use this function to estimate the final context posterior score $s_\phi(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}^c)$. Specifically, the context estimation score is a sum of the single and pairwise vision-language association scores: $\mathbf{x}_j$ and $\mathbf{y}^{c1}$, $[\mathbf{x}_i, \mathbf{x}_j]$ and $\mathbf{y}^{c2}$. Each associate score is an fc output from the input of a normalized feature:

$$\begin{aligned} & \mathbf{m}_j^1 = \mathbf{y}^{c1} \odot \text{fc}(\mathbf{x}_j), \ \mathbf{m}_j^2 = \mathbf{y}^{c2} \odot \text{fc}([\mathbf{x}_i, \mathbf{x}_j]), \\ & \widetilde{\mathbf{m}}_j^1 = \text{L2Norm}(\mathbf{m}_j^1), \ \widetilde{\mathbf{m}}_j^2 = \text{L2Norm}(\mathbf{m}_j^2), \\ & s_\phi(\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}^c) = \text{fc}(\widetilde{\mathbf{m}}_j^1) + \text{fc}(\widetilde{\mathbf{m}}_j^2), \end{aligned} \tag{12}$$

where the element-wise multiplication $\odot$ is an effective way for multimodal features [2]. According to Eq. (4), we can obtain the estimated context z as $\mathbf{z}_i = \sum_j \beta_j \mathbf{x}_j$, where $\beta_j = \text{softmax}_j(s_\phi(\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}^c))$.

Referent Grounding Score: $s_\theta(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}^r)$. After obtaining the context feature $\mathbf{z}_i$, we can use this score function to calculate how likely a candidate RoI $\mathbf{x}_i$ is the referent given the context $\mathbf{z}_i$. This function is similar to Eq. (12).

Context Regularization Score: $s_\omega(\mathbf{z}_i, \mathbf{y}^g) - s_\phi(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}^c)$. As discussed in Eq. (6), this function scores how likely the estimated context feature $\mathbf{z}_i$ is consistent with the content mentioned in the expression. In particular, $s_\omega(\mathbf{z}_i, \mathbf{y}^g)$ is only dependent on single RoI:

$$\mathbf{m}_i{=}\mathbf{y}_i^g \odot \text{fc}(\mathbf{z}_i), \widetilde{\mathbf{m}}_i{=}\text{L2Norm}(\mathbf{m}_i), s_\omega(\mathbf{z}_i, \mathbf{y}_i^g){=}\text{fc}(\mathbf{m}_i). \tag{13}$$

5. Experiment

5.1. Datasets

We used four popular benchmarks for the referring expression grounding task.

RefCOCO [45]. It has 142,210 referring expressions for 50,000 referents (*e.g.*, object instances) in 19,994 images from MSCOCO [18]. The expressions are collected in an interactive way [14]. The dataset is split into train, validation, Test A, and Test B, which has 120,624, 10,834, 5,657 and 5,095 expression-referent pairs, respectively. An image contains multiple people in Test A and multiple objects in Test B.

RefCOCO+ [45]. It has 141,564 expressions for 49,856 referents in 19,992 images from MSCOCO. The difference from RefCOCO is that it only allows appearances but no locations to describe the referents. The split is 120,191, 10,758, 5,726 and 4,889 expression-referent pairs for train, validation, Test A, and Test B respectively.

RefCOCOg [25]. It has 95,010 referring expressions for 49,822 objects in 25,799 images from MSCOCO. Different from RefCOCO and RefCOCO+, this dataset not collected in an interactive way and contains longer sentences containing both appearance and location expressions. The split is 85,474 and 9,536 expression-referent pairs for training and validation. Note that there is no open test split for RefCOCOg, so we used the hyper-parameters cross-validated on RefCOCO and RefCOCO+.

RefCLEF [14]. It contains 20,000 images with annotated image regions. It has some ambiguous (e.g. anywhere) phrases and mistakenly annotated image regions that are not described in the expressions. For fair comparison, we used the split released by [12, 32], *i.e.*, 58,838, 6,333 and 65,193 expression-referent pairs for training, validation and test, respectively.

5.2. Settings and Metrics

We used an English vocabulary of 72,704 words contained in the GloVe pre-trained word vectors [27], which

Table 1. Supervised grounding performances (Acc%) of comparing methods on RefCOCO, RefCOCO+, and RefCOCOg. Note that [46] reports slightly higher accuracies using ensemble models of Listener and Speaker. For fair comparison, we only report their single models.

		State-of-The-Arts						Our Baselines		
Dataset	Split	MMI [25]	NegBag [26]	Attr [19]	CMN [11]	Speaker [46]	Listener [46]	VC w/o reg	VC w/o α	VC
RefCOCO	Test A	71.72	75.6	78.85	75.94	78.95	78.45	75.59	74.03	**78.98**
	Test B	71.09	78.0	78.07	79.57	80.22	80.10	79.69	78.27	**82.39**
RefCOCO+	Test A	58.42	—	61.47	59.29	**64.60**	63.34	60.76	57.61	62.56
	Test B	51.23	—	57.22	59.34	59.62	58.91	60.14	54.37	**62.90**
RefCOCOg	Val	62.14	68.4	69.83	69.30	72.63	72.25	71.05	65.13	**73.98**
RefCOCO(det)	Test A	64.90	58.6	72.08	71.03	72.95	72.95	70.78	70.73	**73.33**
	Test B	54.51	56.4	57.29	65.77	63.43	62.98	65.10	64.63	**67.44**
RefCOCO+(det)	Test A	54.03	—	57.97	54.32	**60.43**	59.61	56.82	53.33	58.40
	Test B	42.81	—	46.20	47.76	48.74	48.44	51.30	46.88	**53.18**
RefCOCOg(det)	Val	45.85	39.5	52.35	57.47	59.51	58.32	60.95	55.72	**62.30**

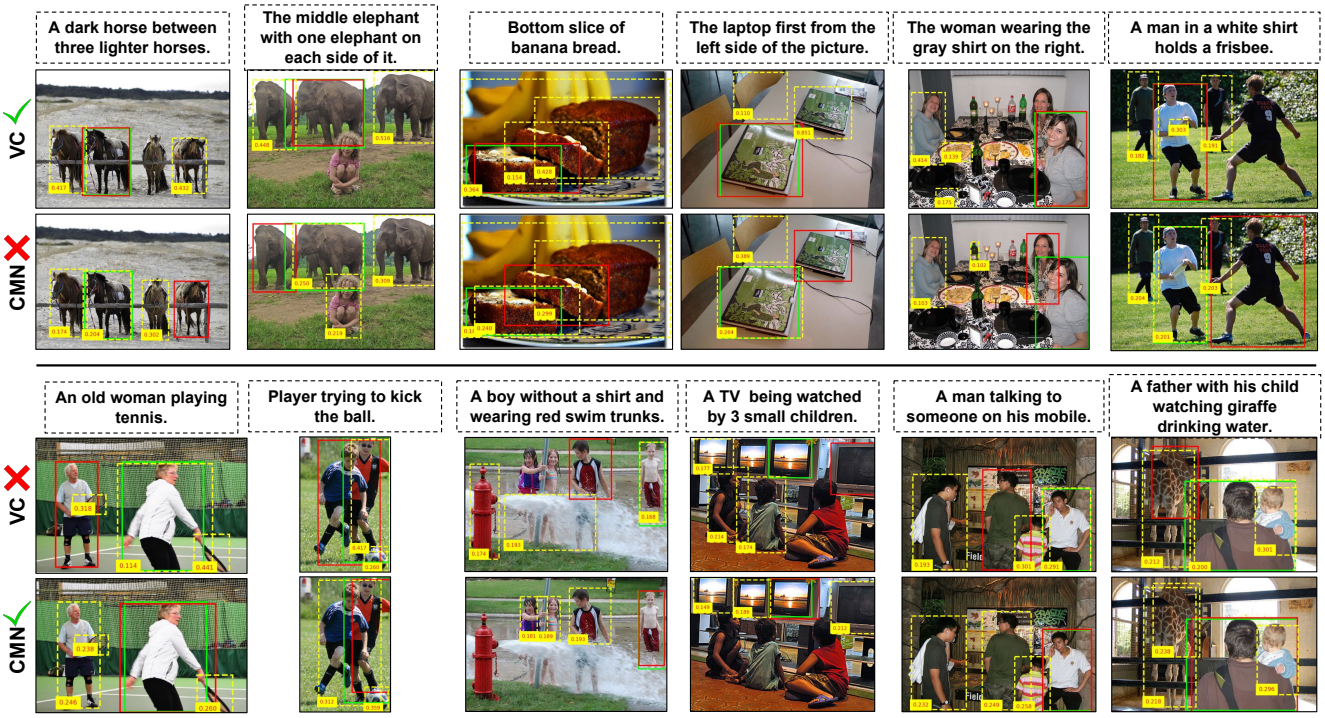

Figure 4. Qualitative results on RefCOCOg (det) showing comparisons between correct (green tick) and wrong referent grounds (red cross) by VC and CMN. The denotations of the bounding box colors are as follows. Solid red: referent ground; solid green: ground truth; dashed yellow: context ground. We only display top 3 context objects with the context ground probability > 0.1. We can observe that VC has more reasonable context localizations than CMN, even in cases when the referent ground of VC fails.

was also used for the initialization of our word vectors. We used a "unk" symbol for the input word of the BLSTM if the word is out of the vocabulary; we set the sentence length to 20 and used "pad" symbol to pad expression sentence < 20. For RoI visual features on RefCOCO, RefCOCO+, and RefCOCOg which have MSCOCO annotated regions with object categories, we used the concatenation of the 4,096-d fc7 output of a VGG-16 based Faster-RCNN network [31] trained on MSCOCO and its corresponding 4,096-d visdiff feature [45]; although RefCLEF regions also have object categories, for fair comparison with [32], we did not use the visdiff feature.

The model training is single-image based, with all referring expressions annotated. We applied SGD of 0.95-momentum with initial learning rate of 0.01, multiplied by 0.1 after every 120,000 iterations, up to 160,000 iterations. Parameters in BILSTM and fc-layers were initialized by Xavier [9] with 0.0005 weight decay. Other settings were default in TensorFlow. Note that our model is trained without bells and whistles, therefore, other optimization tricks

such as batch normalization [13] and GRU [4] are expected to further improve the results reported here. Besides the ground truth annotations, grounding to automatically detected objects is a more practical setting. Therefore, we also evaluated with the SSD-detected bounding boxes [20] on the four datasets provided by [46]. A grounding is considered as correct if the intersection-over-union (IoU) of the top-1 scored region and the ground-truth object is larger than 0.5. The grounding accuracy (a.k.a, P@1) is the fraction of correctly grounded test expressions.

5.3. Evaluations of Supervised Grounding

We compared our variational context model (VC) with state-of-the-art referring expression methods published in recent years, which can be categorized into: 1) generation-comprehension based such as MMI [25], Attr [19], Speaker [46], Listener [46], and SCRC [12]; 2) localization based such as GroundR [32], NegBag [26], CMN [11]. Note that NegBag and CMN are MIL-based models. In particular, we used the author-released code to obtain the results of CMN on RefCLEF, RefCOCO, and RefCOCO+.

From the results on RefCOCO, RefCOCO+, and RefCOCOg in Table 1 and that on RefCLEF in Table 2, we can see that VC achieves the state-of-the-art performance. We believe that the improvement is attributed to the variational Bayesian modeling of context. First, on all datasets, except for the most recent reinforcement learning based [46], VC outperforms all the other sentence generation-comprehension methods that do not model context. Second, compared to VC without the regularization term in Eq. (3) (VC w/o reg), VC can boost the performance by around 2% on all datasets. This demonstrates the effectiveness of the KL divergence for the prevention of the overfitted context estimation.

In particular, we further demonstrate the superiority of VC over the most recent MIL-based method CMN. As illustrated in Figure 4, VC has better context comprehension in both of the language and image regions than CMN. For example, in the top two rows where VC is correct and CMN is wrong, for the grounding in the second column, CMN unnecessarily considers the "girl" as context but the expression only describes using "elephant"; in the last column, CMN misses the key context "frisbee". Even in the failure cases where VC is wrong and CMN is correct, VC still localizes reasonable context. For example, in the fourth column, although CMN grounds the correct TV, but it is based on incorrect context of other TVs; while VC can predict the correct context "children". In addition, we observed that most of the cases that CMN is better than VC involves multiple humans. This demonstrates that VC is better at grounding objects of different categories.

VC is also effective in images with more objects. Figure 5 shows the performances of VC and CMN with various number of bounding boxes. We can observe that VC considerably outperforms CMN over all bounding boxes numbers. Recall that context is the key to distinguish objects of the same category. In particular, on the Test A sets of RefCOCO and RefCOCO+ where the grounding is only about people, *i.e.*, the same object category, the gap between VC and CMN is becoming larger as the box number increases. This demonstrates that MIL is ineffective in modeling context, especially when the number of image regions is large.

Table 2. Performances (Acc%) of supervised and unsupervised methods on RefCLEF.

	Sup.	Sup. (det)	Unsup. (det)
SCRC [12]	72.74	17.93	—
GroundR [32]	—	26.93	10.70
CMN [11]	81.52	28.33	—
VC	**82.43**	**31.13**	14.11
VC w/o α	79.60	27.40	**14.50**

Table 3. Unsupervised grounding performances (Acc%) of comparing methods on RefCOCO, RefCOCO+, and RefCOCOg.

Dataset	Split	VC w/o reg	VC	VC w/o α
RefCOCO	Test A	13.59	17.34	**33.29**
	Test B	21.65	20.98	**30.13**
RefCOCO+	Test A	18.79	23.24	**34.60**
	Test B	24.14	24.91	**31.58**
RefCOCOg	Val	25.14	**33.79**	30.26
RefCOCO(det)	Test A	17.14	20.91	**32.68**
	Test B	22.30	21.77	**27.22**
RefCOCO+(det)	Test A	19.74	25.79	**34.68**
	Test B	24.05	25.54	**28.10**
RefCOCOg(det)	Val	28.14	**33.66**	29.65

5.4. Evaluations of Unsupervised Grounding

We follow the unsupervised setting in GroundR [32]. To our best knowledge, it is the only work on unsupervised referring expression grounding. Note that it is also known as "weakly supervised" detection [48] as there is still image-level ground truth (*i.e.*, the referring expression). Table 2 reports the unsupervised results on the RefCLEF. We can see that VC outperforms the state-of-the-art GroundR, which is a generation-comprehension based method. This demonstrates that using context also helps unsupervised grounding. As there is no published unsupervised results on RefCOCO, RefCOCO+, and RefCOCOg, we only compared our baselines on them in Table 3. We can have the following three key observations which highlight the challenges of unsupervised grounding:

Context Prior. VC w/o reg is the baseline without the KL divergence as a context regularization in Eq. (3). We can see that in most of the cases, VC considerably outperforms VC w/o reg by over 2%, even over 5% on RefCOCO+ (det)

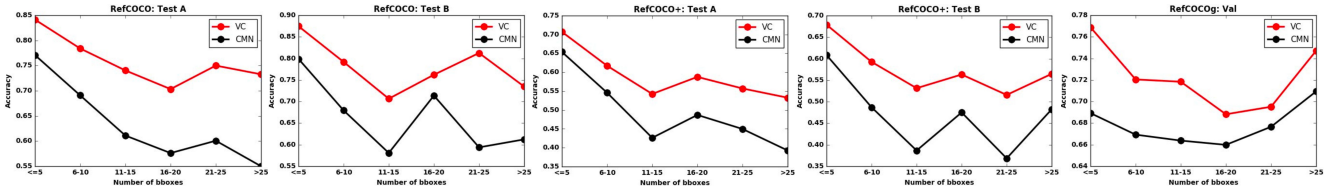

Figure 5. Performances of VC and CMN with different number of object bounding boxes on RefCOCO Test A &B, RefCOCO+ Test A & B, and RefCOCOg Val. Compared to CMN, we can see that VC is more effective in context modeling when the number of objects is large.

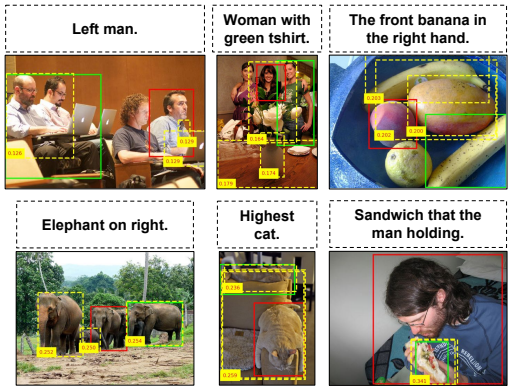

Figure 6. Common failure cases in unsupervised grounding with detected bounding boxes. From left to right: RefCOCO, RefCOCO+, and RefCOCOg. The failure is mainly to the challenging unsupervised relation modeling between referent and context.

and RefCOCOg (det). Note that this improvement is significantly higher than that in supervised setting (*e.g.*, $< 3\%$ as reported in Table 1). The reason is that the context estimation in Eq. (4) would be easier to be stuck in image regions that are irrelevant to the expression in unsupervised setting, therefore, context prior is necessary.

Language Feature. Except on RefCOCOg, we consistently observed the *ineffectiveness* of the cue-specific language feature in unsupervised setting, *i.e.*, VC w/o α outperforms VC in Table 2 and 3. Here α represents the cue-specific word attention. This is contrary to the observation in the supervised setting as listed in Table 1, where VC w/o α is consistently lower than VC. Note that without the cue-specific word attention α in Eq. (10), the language feature is merely the average value of the word embedding vectors in the expression. In this way, VC w/o α does not encode any structural language composition as illustrated in Figure 3, thus, it is better for short expressions. However, when the expression is long in RefCOCOg, discarding the language structure still degrades the performance on RefCOCOg.

Unsupervised Relation Discovery. Although we demonstrated that VC improves the unsupervised grounding by modeling context, we believe that there is still a large space for improving the quality of modeling the context. As the failure examples shown in Figure (6), 1) many context estimations are still out of the scope of the expression, *e.g.*, we may localize the "cup" and "table" as context even though the expression is "woman with green t-shirt"; 2) we

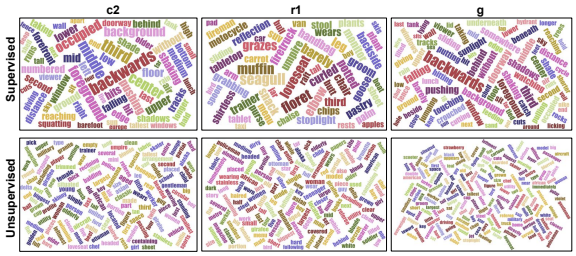

Figure 7. Word cloud visualizations of cue-specific word attention α in Eq. 10 of context-cue (c2), referent-cue (r1), and generic-cue (g) using supervised (top row) and unsupervised training (bottom row) on RefCOCOg. Without supervision, it is difficult to discover meaningful language compositions.

may mistake due to the wrong comprehension of the relations, *e.g.*, "right" as "left", even if the objects belong to the same category, *e.g.*, "elephant". For further investigation, Figure 7 visualizes the cue-specific word attentions in supervised and unsupervised settings. The almost identical word attentions in unsupervised setting reflect the fact that the relation modeling between referent and context is not as successful as in supervised setting. This inspires us to exploit stronger prior knowledge such as language structure [41] and spatial configurations [48, 39].

6. Conclusions

We focused on the task of grounding referring expressions in images and discussed that the key problem is how to model the complex context, which is not effectively resolved by the multiple instance learning framework used in prior works. Towards this challenge, we introduced the Variational Context model, where the variational lower-bound can be interpreted by the reciprocity between the referent and context: given any of which can help to localize the other, and hence is expected to significantly reduce the context complexity in a principled way. We implemented the model using cue-specific language-vision embedding network that can be efficiently trained end-to-end. We validated the effectiveness of this reciprocity by promising supervised and unsupervised experiments on four benchmarks. Moving forward, we are going to 1) incorporate expression language generation in the variational framework, 2) use more structural features of language rather than word attentions, and 3) further investigate the potential of our model in the unsupervised referring expression grounding.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 1

[2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR*, 2015. 2, 5

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. 2015. 4

[4] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 7

[5] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017. 2

[6] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, 2017. 1

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 1997. 2

[8] C. W. Fox and S. J. Roberts. A tutorial on variational bayesian inference. *Artificial intelligence review*, 2012. 2

[9] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *ICAIS*, 2010. 6

[10] D. Golland, P. Liang, and D. Klein. A game-theoretic approach to generating spatial descriptions. In *EMNLP*, 2010. 1, 2

[11] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017. 2, 3, 4, 6, 7

[12] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *CVPR*, 2016. 2, 5, 7

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 7

[14] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2, 5

[15] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2, 3

[16] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, and M. Zhou. Visual question generation as dual task of visual question answering. In *CVPR*, 2018. 1

[17] Y. Li, W. Ouyang, and X. Wang. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*, 2017. 2

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[19] J. Liu, L. Wang, and M.-H. Yang. Referring expression generation and comprehension via attributes. In *ICCV*, 2017. 2, 6, 7

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 4, 7

[21] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 2

[22] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 4

[23] R. Luo and G. Shakhnarovich. Comprehension-guided referring expressions. In *CVPR*, 2017. 2

[24] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow. Adversarial autoencoders. In *ICLR Workshop*, 2016. 3

[25] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1, 2, 3, 5, 6, 7

[26] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2, 3, 6, 7

[27] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 5

[28] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik. Phrase localization and visual relationship detection with comprehensive linguistic cues. In *ICCV*, 2017. 1, 2

[29] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2

[30] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 1

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 6

[32] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 2, 5, 6, 7

[33] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *TSP*, 1997. 4

[34] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Workshop on Vision and Language*, 2015. 1

[35] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015. 3

[36] Q. Sun, B. Schiele, and M. Fritz. A domain based approach to social relation recognition. In *CVPR*, 2017. 1

[37] J. A. Thomas. *Meaning in interaction: An introduction to pragmatics*. Routledge, 2014. 2

[38] J. Thomason, J. Sinapov, and R. Mooney. Guiding interaction behaviors for multi-modal grounded language learning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 20–24, 2017. 1

[39] Y. Wei, J. Feng, X. Liang, C. Ming-Ming, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017. 8

[40] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning*. 1992. 3

[41] F. Xiao, L. Sigal, and Y.-J. Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, 2017. 8

[42] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[43] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*, 2016. 2

[44] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016. 2

[45] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2, 4, 5, 6

[46] L. Yu, H. Tan, M. Bansal, and T. L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *ICCV*, 2017. 2, 6, 7

[47] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017. 1, 2

[48] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *ICCV*, 2017. 2, 7, 8

[49] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, 2017. 1

[50] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 4