Modeling Relationships in Referential Expressions with Compositional Modular Networks

Ronghang Hu[1] Marcus Rohrbach[1] Jacob Andreas[1] Trevor Darrell[1] Kate Saenko[2]
[1]University of California, Berkeley [2]Boston University
{ronghang,rohrbach,jda,trevor}@eecs.berkeley.edu, saenko@bu.edu

Abstract

People often refer to entities in an image in terms of their relationships with other entities. For example, the black cat sitting under the table *refers to both a* black cat *entity and its relationship with another* table *entity. Understanding these relationships is essential for interpreting and grounding such natural language expressions. Most prior work focuses on either grounding entire referential expressions holistically to one region, or localizing relationships based on a fixed set of categories. In this paper we instead present a modular deep architecture capable of analyzing referential expressions into their component parts, identifying entities and relationships mentioned in the input expression and grounding them all in the scene. We call this approach Compositional Modular Networks (CMNs): a novel architecture that learns linguistic analysis and visual inference end-to-end. Our approach is built around two types of neural modules that inspect local regions and pairwise interactions between regions. We evaluate CMNs on multiple referential expression datasets, outperforming state-of-the-art approaches on all tasks.*

1. Introduction

Great progress has been made on object detection, the task of localizing visual entities belonging to a pre-defined set of categories [8, 23, 22, 6, 16]. But the more general and challenging task of localizing entities based on arbitrary natural language expressions remains far from solved. This task, sometimes known as *grounding* or *referential expression comprehension*, has been explored by recent work in both computer vision and natural language processing [19, 10, 24]. Given an image and a natural language expression referring to a visual entity, such as *the young man wearing green shirt and riding a black bicycle*, these approaches localize the image region corresponding to the entity that the expression refers to with a bounding box.

Referential expressions often describe relationships between multiple entities in an image. In Figure 1, the expression *the woman holding a grey umbrella* describes a *woman* entity that participates in a *holding* relationship with a *grey umbrella* entity. Because there are multiple women in the image, resolving this referential expression requires both finding a bounding box that contains a person, and ensuring that this bounding box relates in the right way to other objects in the scene. Previous work on grounding referential expressions either (1) treats referential expressions holistically, thus failing to model explicit correspondence between textual components and visual entities in the image [19, 10, 24, 32, 20], or else (2) relies on a fixed set of

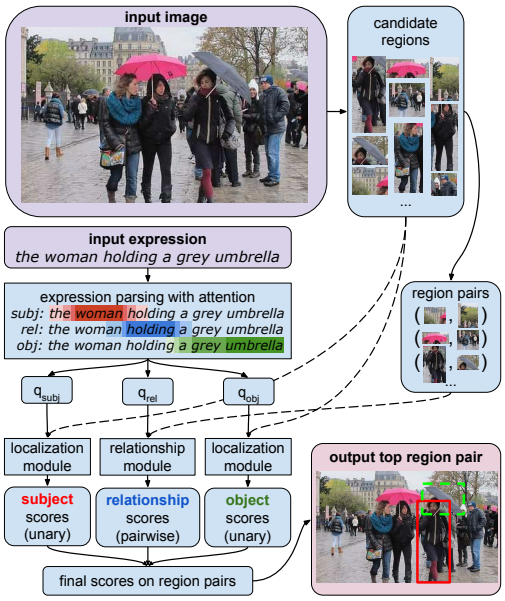


Figure 1. Given an image and an expression, we learn to parse the expression into vector representation of subject q_{subj}, relationship q_{rel} and object q_{obj} with attention, and align these textual components to image regions with two types of modules. The localization module outputs scores over each individual region while the relationship module produces scores over region pairs. These outputs are integrated into final scores over region pairs, producing the top region pair as grounding result. (Best viewed in color.)

entity and relationship categories defined *a priori* [17].

In this paper, we present a joint approach that explicitly models the compositional linguistic structure of referential expressions and their groundings, but which nonetheless supports interpretation of arbitrary language. We focus on referential expressions involving inter-object relationships that can be represented as a subject entity, a relationship and an object entity. We propose Compositional Modular Networks (CMNs), an end-to-end trained model that learns language representation and image region localization jointly as shown in Figure 1. Our model differentiably parses the referential expression into a subject, relationship and object with three soft attention maps, and aligns the extracted textual representations with image regions using a modular neural architecture. There are two types of modules in our model, one used for localizing specific textual components by outputting unary scores over regions for that component, and one for determining the relationship between two pairs of bounding boxes by outputting pairwise scores over region-region pairs. We evaluate our model on multiple datasets, and show that our model outperforms both natural baselines and previous work.

2. Related work

Grounding referential expressions. The problem of grounding referential expressions can be naturally formulated as a retrieval problem over image regions [19, 10, 24, 7, 32, 20]. First, a set of candidate regions are extracted (*e.g.* via object proposal methods like [28, 4, 12, 35]). Next, each candidate region is scored by a model with respect to the query expression, returning the highest scoring candidate as the grounding result. In [19, 10], each region is scored based on its local visual features and some global contextual features from the whole image. However, local visual features and global contextual from the whole image are often insufficient to determine whether a region matches an expression, as relationships with other regions in the image must also be considered. Two recent methods [32, 20] go beyond local visual features in a single region, and consider multiple regions at the same time. [32] adds contextual feature extracted from other regions in the image, and [20] proposes a model that grounds a referential expression into a pair of regions. All these methods represent language holistically using a recurrent neural network: either generatively, by predicting a distribution over referential expressions [19, 10, 32, 20], or discriminatively, by encoding expressions into a vector representation [24, 7]. This makes it difficult to learn explicit correspondences between the components in the textual expression and entities in the image. In this work, we learn to parse the language expression into textual components in instead of treating it as a whole, and align these components with image regions end-to-end.

Handling inter-object relationships. Recently work by [17] trains detectors based on RCNN [8] and uses a linguistic prior to detect visual relationships. However, this work relies on fixed, predefined categories for subjects, relations, and objects, treating entities like "bicycle" and relationships like and "riding" as discrete classes. Instead of building upon a fixed inventory of classes, our model handles relationships specified by arbitrary natural language phrases, and jointly learns expression parsing and visual entity localization. Although [14] also learns language parsing and perception, it is directly based on logic (λ-calculus) and requires additional classifiers trained for each predicate class. Aside from localizing relationship expressions, [30] generates image descriptions using a recurrent network with attention over image feature grids, and [25, 31] learns to extract visual relation knowledge from images.

Compositional structure with modules. Neural Module Networks [3] address visual question answering by decomposing the questions into textual components and dynamically assembling a specific network architecture for the question from a few network modules based on the textual components. However, this method relies on an external language parser for textual analysis instead of end-to-end learned language representation, and is not directly applicable to the task of grounding referential expressions into bounding boxes, since it does not explicitly output bounding boxes as results. Recently, [2] improves over [3] by learning to re-rank parsing outputs from the external parser, but it is still not end-to-end learned since the parser is fixed and not optimized for the task. Inspired by [3], our model also uses a modular structure, but learns the language representation end-to-end from words.

3. Our model

We propose Compositional Modular Networks (CMNs) to localize visual entities described by a query referential expression. Our model is compositional in the sense that it localizes a referential expression by grounding the components in the expressions and exploiting their interactions, in accordance with the principle of compositionality of natural language – the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them [29]. Our model works in a retrieval setting: given an image I, a referential expression Q as query and a set of candidate region bounding boxes $B = \{b_i\}$ for the image I (*e.g.* extracted through object proposal methods), our model outputs a score for each bounding box b_i, and returns the bounding box with the highest score as grounding (localization) result. Unlike state-of-the-art methods [24, 7], the scores for each region bounding box $b_i \in B$ are not predicted only from the local feature of b_i, but also based on other regions in the image. In our model, we focus on the relationships in referential expressions that can be represented as a 3-component triplet `(subject,`

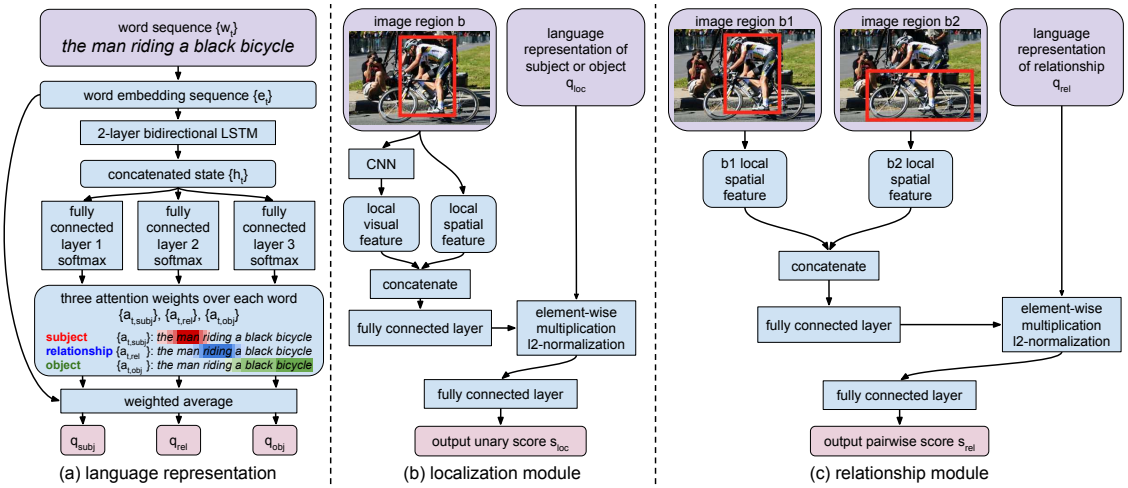


Figure 2. Detailed illustration of our model. (a) Our model learns to parse an expression into subject, relationship and object with attention for language representation (Sec. 3.1). (b) The localization module matches subject or object with each image region and returns a unary score (Sec. 3.2). (c) The relationship module matches a relationship with a pair of regions and returns a pairwise score (Sec. 3.3).

`relationship, object)`, and learn to parse the expressions into these components with attention. For example, *a young man wearing a blue shirt* can be parsed as the triplet (*a young man*, *wearing*, *a blue shirt*). The score of a region is determined by simultaneously looking at whether it matches the description of the subject entity and whether it matches the relationship with another interacting object entity mentioned in the expression.

Our model handles such inter-object relationships by looking at pairs of regions (b_i, b_j). For referential expressions like "the red apple on top of the bookshelf", we want to find a region pair (b_i, b_j) such that b_i matches the subject entity "red apple" and b_j matches the object entity "bookshelf" and the configuration of (b_i, b_j) matches the relationship "on top of". To achieve this goal, our model is based on a compositional modular structure, composed of two modules assembled in a pipeline for different sub-tasks: one localization module $f_{loc}(\cdot, q_{loc}; \Theta_{loc})$ for deciding whether a region matches the subject or object in the expression, where q_{loc} is the textual vector representation of the subject component "red apple" or the object component "bookshelf", and one relationship module $f_{rel}(\cdot, \cdot, q_{rel}; \Theta_{rel})$ for deciding whether a pair of regions matches the relationship described in the expression represented by q_{rel}, the textual vector representation of the relationship "on top of". The representations q_{subj}, q_{rel} and q_{obj} are learned jointly in our model in Sec. 3.1.

We define the pairwise score $s_{pair}(b_i, b_j)$ over a pair of image regions (b_i, b_j) matching an input referential expression Q as the sum of three components:

$$\begin{aligned} s_{pair}(b_i, b_j) = & f_{loc}(b_i, q_{subj}; \Theta_{loc}) \\ & + f_{loc}(b_j, q_{obj}; \Theta_{loc}) \\ & + f_{rel}(b_i, b_j, q_{rel}; \Theta_{rel}), \end{aligned} \tag{1}$$

where q_{subj}, q_{obj} and q_{rel} are vector representations of subject, relationship and object, respectively.

For inference, we define the final subject unary score $s_{subj}(b_i)$ of a bounding of b_i corresponding to the subject (*e.g.* "the red apple" in "the red apple on top of the bookshelf") as the score of the best possible pair (b_i, b_j) that matches the entire expression:

$$s_{subj}(b_i) \triangleq \max_{b_j \in B} s_{pair}(b_i, b_j). \tag{2}$$

The subject is ultimately grounded (localized) to the highest scoring region as $b^*_{subj} = \arg\max_{b_i \in B}(s_{subj}(b_i))$.

3.1. Expression parsing with attention

Given a referential expression Q like *the tall woman carrying a red bag*, how can we decide which substrings corresponds to the subject, the relationship, and the object, and extract three vector representations q_{subj}, q_{rel} and q_{obj} corresponding to these three components? One possible approach is to use an external language parser to parse the referential expression into the triplet format `(subject, relationship, object)` and then process each component with an encoder (*e.g.* a recurrent neural network) to extract q_{subj}, q_{rel} and q_{obj}. However, the formal representations of language produced by syntactic parsers do not always correspond to intuitive visual representations. As a simple example, *the apple on top of the bookshelf* is analyzed [33] as having a subject phrase *the apple*, a relationship *on*, and an object phrase *top of the bookshelf*, when in fact the visually salient objects are simply the apple and the bookshelf, while the complete expression *on top of* describes the relationship between them.

Therefore, in this work we learn to decompose the input expression Q into the above 3 components, and generate vector representations q_{subj}, q_{rel} and q_{obj} from Q through a soft attention mechanism over the word sequence, as shown

in Figure 2 (a). For a referential expression Q that is a sequence of T words $\{w_t\}_{t=1}^T$, we first embed each word w_t to a vector e_t using GloVe [21], and then scan through the word embedding sequence $\{e_t\}_{t=1}^T$ with a 2-layer bi-directional LSTM network [26]. The first layer takes as input the sequence $\{e_t\}$ and outputs a forward hidden state $h_t^{(1,fw)}$ and a backward hidden state $h_t^{(1,bw)}$ at each time step, which are concatenated into $h_t^{(1)}$. The second layer then takes the first layer's output sequence $\{h_t^{(1)}\}$ as input and outputs forward and backward hidden states $h_t^{(2,fw)}$ and $h_t^{(2,bw)}$ at each time step. All the hidden states in the first layer and second layer are concatenated into a single vector $h_t = \left[h_t^{(1,fw)} \;\; h_t^{(1,bw)} \;\; h_t^{(2,fw)} \;\; h_t^{(2,bw)}\right]$.

The concatenated state h_t contains information from word w_t itself and also context from words before and after w_t. Then the attention weights $a_{t,subj}$, $a_{t,rel}$ and $a_{t,obj}$ for subject, relationship, object over each word w_t are obtained by three linear predictions over h_t followed by a softmax as

$$a_{t,subj} = \exp\left(\beta_{subj}^T h_t\right) / \sum_{\tau=1}^{T} \exp\left(\beta_{subj}^T h_\tau\right) \quad (3)$$

$$a_{t,rel} = \exp\left(\beta_{rel}^T h_t\right) / \sum_{\tau=1}^{T} \exp\left(\beta_{rel}^T h_\tau\right) \quad (4)$$

$$a_{t,obj} = \exp\left(\beta_{obj}^T h_t\right) / \sum_{\tau=1}^{T} \exp\left(\beta_{obj}^T h_\tau\right) \quad (5)$$

and the language representations of the subject q_{subj}, relationship q_{rel} and object q_{obj} are extracted as weighed average of word embedding vectors $\{e_t\}$ with attention weights as $q_{subj} = \sum_{t=1}^T a_{t,subj} e_t$, and $q_{rel} = \sum_{t=1}^T a_{t,rel} e_t$ and $q_{obj} = \sum_{t=1}^T a_{t,obj} e_t$.

3.2. Localization module

As shown in Figure 2 (b), the localization module f_{loc} outputs a score $s_{loc} = f_{loc}(b, q_{loc}; \Theta_{loc})$ representing how likely a region bounding box b matches q_{loc}, which is either the subject textual vector q_{subj} or object textual vector q_{obj}.

This module takes the local visual feature x_{vis} and spatial feature $x_{spatial}$ of image region b. We extract visual feature x_v from image region b using a convolutional neural network [27], and extract a 5-dimensional spatial feature $x_s = [\frac{x_{min}}{W_I}, \frac{y_{min}}{H_I}, \frac{x_{max}}{W_I}, \frac{y_{max}}{H_I}, \frac{S_b}{S_I}]$ from b using the same representation as in [19], where $[x_{min}, y_{min}, x_{max}, y_{max}]$ and S_b are bounding box coordinates and area of b, and W_I, H_I and S_I are width, height and area of the image I. Then, x_v and x_s are concatenated into a vector $x_{v,s} = [x_v \; x_s]$ as representation of region b.

Since element-wise multiplication is shown to be a powerful way to combine representations from different modalities [5], we adopt it here to obtain a joint vision and language representation. In our implementation, $x_{v,s}$ is first embedded to a new vector $\tilde{x}_{v,s}$ that has the same dimension as q_{loc} (which is either q_{subj} or q_{obj}) through a linear transform, and then element-wise multiplied with q_{loc} to obtain a vector z_{loc}, which is L2-normalized into $\hat{z}_{loc}$ to obtain a more robust representation, as follows:

$$\tilde{x}_{v,s} = W_{v,s} x_{v,s} + b_{v,s} \quad (6)$$

$$z_{loc} = \tilde{x}_{v,s} \odot q_{loc} \quad (7)$$

$$\hat{z}_{loc} = z_{loc} / \|z_{loc}\|_2 \quad (8)$$

where $\odot$ is element-wise multiplication between two vectors. Then the score s_{loc} is predicted linearly from $\hat{z}_{loc}$ as

$$s_{loc} = w_{loc}^T \hat{z}_{loc} + b_{loc}. \quad (9)$$

The parameters in Θ_{loc} are $(W_{v,s}, b_{v,s}, w_{loc}, b_{loc})$.

3.3. Relationship module

As shown in Figure 2 (c), the relationship module f_{rel} outputs a score $s_{rel} = f_{rel}(b_1, b_2, q_{rel}; \Theta_{rel})$ representing how likely a pair of region bounding boxes (b_1, b_2) matches q_{rel}, the representation of relationship in the expression.

In our implementation, we use the spatial features x_{s1} and x_{s2} of the two regions b_1 and b_2 extracted in the same way as in localization module (we empirically find that adding visual features of b_1 and b_2 leads to no noticeable performance boost while slowing training significantly). Then x_{s1} and x_{s2} are concatenated as $x_{s1,s2} = [x_{s1} \; x_{s2}]$, and then processed in a similar way as in localization module to obtain s_{rel}, as shown below:

$$\tilde{x}_{s1,s2} = W_{s1,s2} x_{s1,s2} + b_{s1,s2} \quad (10)$$

$$z_{rel} = \tilde{x}_{s1,s2} \odot q_{rel} \quad (11)$$

$$\hat{z}_{rel} = z_{rel} / \|z_{rel}\|_2 \quad (12)$$

$$s_{rel} = w_{rel}^T \hat{z}_{rel} + b_{rel}. \quad (13)$$

The parameters in Θ_{rel} are $(W_{s1,s2}, b_{s1,s2}, w_{rel}, b_{rel})$.

3.4. End-to-end learning

During training, for an image I, a referential expression Q and a set of candidate regions B extracted from I, if the ground-truth regions b_{subj_gt} of the subject entity and b_{obj_gt} of the object entity are both available, then we can optimize the pairwise score s_{pair} in Eqn. 1 with strong supervision using softmax loss $Loss_{strong}$.

$$Loss_{strong} = -\log\left(\frac{\exp\left(s_{pair}(b_{subj\_gt}, b_{obj\_gt})\right)}{\sum_{(b_i,b_j)\in B\times B} \exp\left(s_{pair}(b_i, b_j)\right)}\right) \quad (14)$$

However, it is often hard to obtain ground-truth regions for both subject entity and object entity. For referential expressions like "a red vase on top of the table", often there is only

Method	Accuracy
baseline (loc module)	46.27%
our full model	99.99%

Table 1. Accuracy of our model and the baseline on the synthetic shape dataset. See Sec. 4.1 for details.

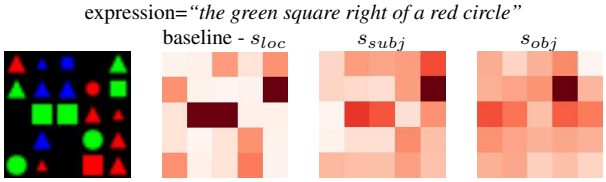


Figure 3. For the image in (a) and the expression "the green square right of a red circle", (b) baseline scores on each location on the 5 by 5 grid using localization module only (darker is higher), and (c, d) scores s_{subj} and s_{obj} using our full model. s_{subj} is highest on the exact green square that is on the right of a red circle, and s_{obj} is highest on this red circle.

a ground-truth bounding box annotation b_1 for the subject (vase) in the expression, but no bounding box annotation b_2 for the object (table), so one cannot directly optimize the pairwise score $s_{pair}(b_1, b_2)$. To address this issue, we treat the object region b_2 as a latent variable, and optimize the unary score $s_{subj}(b_1)$ in Eqn. 2. Since $s_{subj}(b_1)$ is obtained by maximizing over all possible region $b_2 \in B$ in $s_{pair}(b_1, b_2)$, this can be regarded as a weakly supervised Multiple Instance Learning (MIL) approach similar to [20]. The unary score s_{subj} can be optimized with weak supervision using softmax loss $Loss_{weak}$.

$$Loss_{weak} = -\log\left(\frac{\exp\left(s_{subj}(b_{subj\_gt})\right)}{\sum_{b_i \in B}\exp\left(s_{subj}(b_i)\right)}\right) \quad (15)$$

The whole system is trained end-to-end with backpropagation, and parameters in localization module, relationship module, language representation and visual feature extraction (convolutional neural network) are jointly optimized. Our model is implemented using TensorFlow [1] and our code is available at `http://ronghanghu.com/cmn`.

4. Experiments

We first evaluate our model on a synthetic dataset to verify its ability to handle inter-object relationships in referential expressions. Next we apply our method to real images and expressions in the Visual Genome dataset [13] and Google-Ref dataset [19]. Since the task of answering pointing questions in visual question answering is similar to grounding referential expressions, we also evaluate our model on the pointing questions in the Visual-7W dataset [34].

4.1. Analysis on a synthetic dataset

Inspired by [3], we first perform a simulation experiment on a synthetic shape dataset. The dataset consists of 30000 images with simple circles, squares and triangles of different sizes and colors on a 5 by 5 grid, and referential expressions constructed using a template of the form `[subj] [relationship] [obj]`, where `[subj]` and `[obj]` involve both shape classes and attributes and `[relationship]` is some spatial relationships such as "above". The task is to localize the corresponding shape region described by the expression on the 5 by 5 grid. Figure 3 (a) shows an example in this dataset with the synthetic expression "the green square right of a red circle". In the synthesizing procedure, we make sure that the shape region being referred to cannot be inferred simply from `[subj]` as there will be multiple matching regions, and the relationship with another region described by `[obj]` has to be taken into consideration.

On this dataset, we train our model with weak supervision by Eqn. 15 using the ground-truth subject region b_{subj_gt} of the subject shape described in the expression. Here the candidate region set B are the 25 possible locations on the 5 by 5 grid, and visual features are extracted from the corresponding cropped image region with a VGG-16 network [27] pretrained on ImageNET classification. As a comparison, we also train a baseline model using only the localization module, with a softmax loss on its output s_{loc} in Eqn. 9 over all 25 locations on the grid, and language representation q_{loc} obtained by scanning through the word embedding sequence with a single LSTM network and taking the hidden state at the last time step same as in [24, 9]. This baseline method resembles the supervised version of GroundeR [24], and the main difference between this baseline and our model is that the baseline only looks at a region's appearance and spatial property but ignores pairwise relationship with other regions.

We evaluate with the accuracy on whether the predicted subject region b^*_{subj} matches the ground-truth region b_{subj_gt}. Table 1 shows the results on this dataset, where our model trained with weak supervision (the same as the supervision given to baseline) achieves nearly perfect accuracy—significantly outperforming the baseline using a localization module only. Figure 3 shows an example, where the baseline can localize green squares but fails to distinguish the exact green square right of a red circle, while our model successfully finds the subject-object pair, although it has never seen the ground-truth location for the object entity during training.

4.2. Localizing relationships in Visual Genome

We also evaluate our method on the Visual Genome dataset [13], which contains relationship expressions annotated over pairs of objects, such as "computer on top of ta-

Method	training supervision	P@1-subj	P@1-pair
baseline	subject-GT	41.20%	-
baseline	subject-object-GT	-	23.37%
our full model	subject-GT	43.81%	26.56%
our full model	subject-object-GT	**44.24%**	**28.52%**

Table 2. Performance of our model on relationship expressions in Visual Genome dataset. See Sec. 4.2 for details.

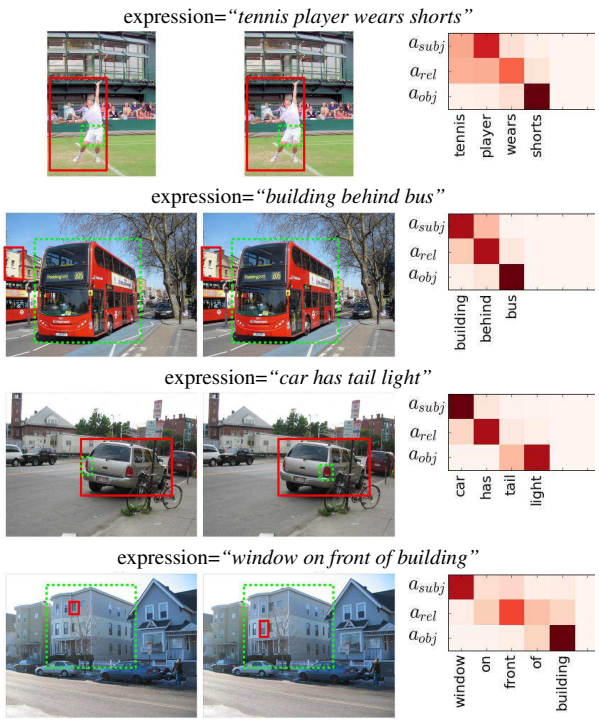
ble" and "person wearing shirt".

On the relationship annotations in Visual Genome, given an image and an expression like "man wearing hat", we evaluate our method in two test scenarios: retrieving the *subject* region ("man") and retrieving the *subject-object pair* (both "man" and "hat"). In our experiment, we take the bounding boxes of all the annotated entities in each image (around 35 per image) as candidate region set B at both training and test time, and extract visual features for each region from fc7 output of a Faster-RCNN VGG-16 network [23] pretrained on MSCOCO detection dataset [15]. We use the same training, validation and test split as in [11].

Since there are ground-truth annotations for both subject region and object region in this dataset, we experiment with two training supervision settings: (1) *weak supervision* by only providing the ground-truth region of the subject entity at training time (**subject-GT** in Table 2) and optimizing unary subject score s_{subj} with Eqn. 15 and (2) *strong supervision* by providing the ground-truth region pair of both subject and object entities at training time (**subject-object-GT** in Table 2) and optimizing pairwise score s_{pair} with Eqn. 14.

Similar to the experiment on the synthetic dataset in Sec. 4.1, we also train a baseline model that only looks at local appearance and spatial properties but ignores pairwise relationships. For the first evaluation scenario of retrieving the subject region, we train a baseline model using a localization module only by optimizing its output s_{loc} for ground-truth subject region with softmax loss (the same training supervision as subject-GT). For the second scenario of retrieving the subject-object pair, we train two such baseline models optimized with subject ground-truth and object ground-truth respectively, to localize of the subject region and object region separately with each model and at test time combine the predicted subject region and predicted object region from each model be the subject-object pair (same training supervision as subject-object-GT).

We evaluate with top-1 precision (P@1), which is the percentage of test instances where the top scoring prediction matches the ground-truth in each image (P@1-subj for predicted subject regions matching subject ground-truth in the first scenario, and P@1-pair for predicted subject and object regions both matching the ground-truth in the second scenario). The results are summarized in Table 2, where it can be seen that our full model outperforms the baseline



(a) ground-truth (b) our prediction (c) attention weights

Figure 4. Visualization of grounded relationship expressions in the Visual Genome dataset, trained with weak supervision (subject-GT). (a, b) ground-truth region pairs and our predicted region pairs respectively (subject in red solid box and object in green dashed box). (c) attention weights in Eqn. 3–5 for subject, relationship and object (darker is higher).

using only localization modules in both evaluation scenarios. Note that in the second evaluation scenario of retrieving subject-object pairs, our weakly supervised model still outperforms the baseline trained with strong supervision.

Figure 4 shows some examples of our model trained with weak supervision (subject-GT) and attention weights in Eqn. 3–5. It can be seen that even with weak supervision, our model still generates reasonable attention weights over words for subject, relationship and object.

4.3. Grounding referential expressions in images

We apply our model to the Google-Ref dataset [19], a benchmark dataset for grounding referential expressions. As this dataset does not explicitly contain subject-object pair annotation for the referential expressions, we train our model with weak supervision (Eqn. 15) by optimizing the subject score s_{subj} using the expression-level region ground-truth. The candidate bounding box set B at both training and test time are all the annotated entities in the image (which is the "Ground-Truth" evaluation setting in [19]). As in Sec. 4.2, fc7 output of a MSCOCO-pretrained Faster-RCNN VGG-16 network is used for visual feature extraction. Similar to Sec. 4.1, we also train a GroundeR-

Method	P@1
Mao *et al*. [19]	60.7%
Yu *et al*. [32]	64.0%
Nagaraja *et al*. [20]	68.4%
baseline (loc module)	66.5%
our model (w/ external parser)	53.5%
our full model	**69.3%**

Table 3. Top-1 precision of our model and previous methods on Google-Ref dataset. See Sec. 4.3 for details.

like [24] baseline model with localization module which looks only at a region's local features.

In addition, instead of learning a linguistic analysis end-to-end as in Sec. 3.1, we also experiment with parsing the expression using the Stanford Parser [33, 18]. An expression is parsed into subject, relationship and object component according to the constituency tree, and the components are encoded into vectors q_{subj}, q_{rel} and q_{obj} using three separate LSTM encoders, similar to the baseline and [24].

Following [19], we evaluate on this dataset using the top-1 precision (P@1) metric, which is the fraction of the highest scoring subject region matching the ground-truth for the expression. Table 3 shows the performance of our model, baseline model and previous work. Note that all the methods are trained with the same weak supervision (only a ground-truth subject region). It can be seen that by incorporating inter-object relationships, our full model outperforms the baseline using only localization modules, and works better than previous state-of-the-art methods.

Additionally, replacing the learned expression parsing and language representation in Sec. 3.1 with an external parser ("our model w/ external parser" in Table 3) leads to a significant performance drop. We find that this is mainly because existing parsers are not specifically tuned for the referring expression task—as noted in Sec. 3.1, expressions like *chair on the left of the table* are parsed as (*chair*, *on*, *the left of the table*) rather than the desired triplet (*chair*, *on the left of*, *the table*). In our full model, the language representation is end-to-end optimized with other parts, while it is hard to jointly optimize an external language parser like [33] for this task.

Figure 5 shows some example results on this dataset. It can be seen that although weakly supervised, our model not only grounds the subject region correctly (solid box), but also finds reasonable regions (dashed box) for the object entity.

4.4. Answering pointing questions in Visual-7W

Finally, we evaluate our method on the multiple choice pointing questions (*i.e*. "which" questions) in visual question answering on the Visual-7W dataset [34]. Given an image and a question like "which tomato slice is under the knife", the task is to select the corresponding region from a few choice regions (4 choices in this dataset) as answer. Since this task is closely related to grounding referential expressions, our model can be trained in the same way as in Sec. 4.3 to score each choice region using subject score s_{subj} and pick the highest scoring choice as answer.

Method	Accuracy
Zhu *et al*. [34]	56.10%
baseline (loc module)	71.61%
our model (w/ external parser)	61.66%
our full model	**72.53%**

Table 4. Accuracy of our model and previous methods on the pointing questions in Visual-7W dataset. See Sec. 4.4 for details.

As before, we train our model with weak supervision through Eqn. 15 and use a MSCOCO-pretrained Faster-RCNN VGG-16 network for visual feature extraction. Here we use two different candidate bounding box sets B_{subj} and B_{obj} of the subject regions (the choices) and the object regions, where B_{subj} is the 4 choice bounding boxes, and B_{obj} is the set of 300 proposal bounding boxes extracted using RPN in Faster-RCNN [23]. Similar to Sec. 4.3, we also train a baseline model using only a localization module to score each choice based only on its local appearance and spatial properties, and a truncated model that uses the Stanford parser [33, 18] for expression parsing and language representation.

The results are shown in Table 4. It can be seen that our full model outperforms the baseline and the truncated model with an external parser, and achieves much higher accuracy than previous work [34]. Figure 6 shows some question answering examples on this dataset.

5. Conclusion

We have proposed Compositional Modular Networks, a novel end-to-end trainable model for handling relationships in referential expressions. Our model learns to parse input expressions with soft attention, and incorporates two types of modules that consider a region's local features and pairwise interaction between regions respectively. The model induces intuitive linguistic and visual analyses of referential expressions from only weak supervision, and experimental results demonstrate that our approach outperforms both natural baselines and state-of-the-art methods on multiple datasets.

Acknowledgements

This work was supported by DARPA, AFRL, DoD MURI award N000141110688, NSF awards IIS-1427425, IIS-1212798 and IIS-1212928, NGA and the Berkeley Artificial Intelligence Research (BAIR) Lab. Jacob Andreas is supported by a Facebook graduate fellowship and a Huawei / Berkeley AI fellowship.

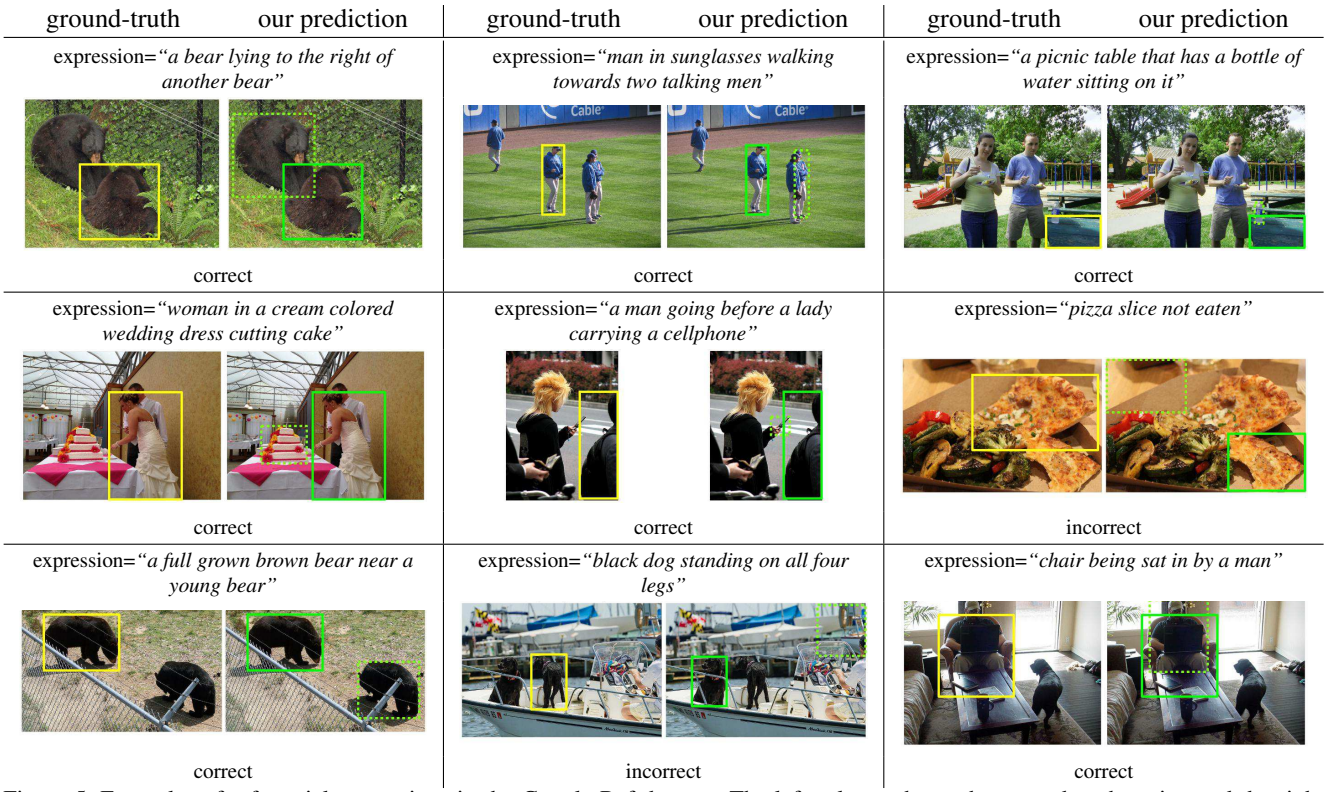


Figure 5. Examples of referential expressions in the Google-Ref dataset. The left column shows the ground-truth region and the right column shows the grounded subject region (our prediction) in solid box and the grounded object region in dashed box. A prediction is labeled as correct if the predicted subject region matches the ground-truth region.

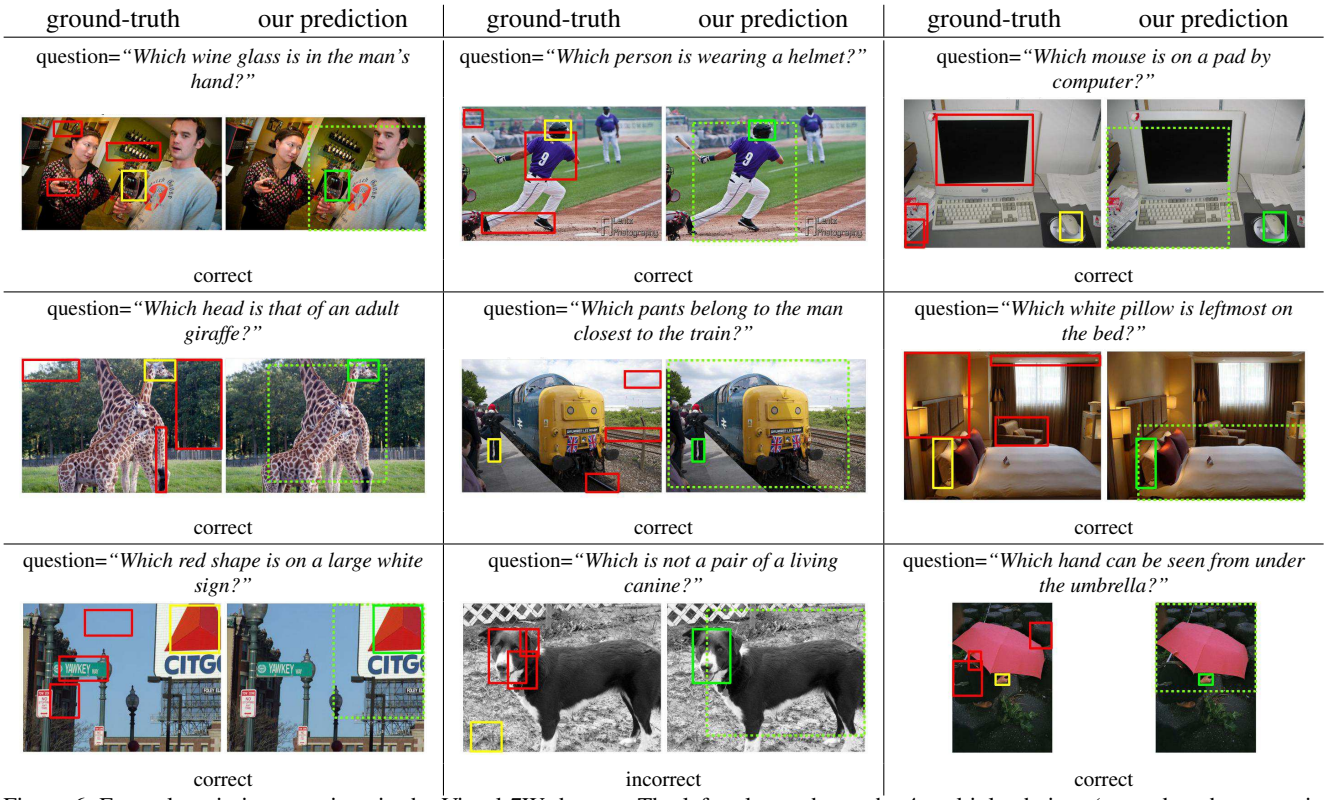


Figure 6. Example pointing questions in the Visual-7W dataset. The left column shows the 4 multiple choices (ground-truth answer in yellow) and the right column shows the grounded subject region (predicted answer) in solid box and the grounded object region in dashed box. A prediction is labeled as correct if the predicted subject region matches the ground-truth region.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems. *arXiv:1603.04467*, 2016. 5

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016. 2

[3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 5

[4] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[5] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 4

[6] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 1

[7] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 2

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[9] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 5

[10] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[11] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[12] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2

[13] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 5

[14] J. Krishnamurthy and T. Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013. 2

[15] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 6

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[17] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[18] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. 7

[19] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 4, 5, 6, 7

[20] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 5, 7

[21] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 4

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[23] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 1, 6, 7

[24] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 5, 7

[25] F. Sadeghi, S. K. Kumar Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[26] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 4

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings*

of the International Conference on Learning Representations (ICLR), 2015. 4, 5

[28] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 2

[29] M. Werning, W. Hinzen, and E. Machery. *The Oxford handbook of compositionality*. Oxford University Press, 2012. 2

[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015. 2

[31] M. Yatskar, V. Ordonez, and A. Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of NAACL-HLT*, 2016. 2

[32] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 7

[33] M. Zhu, Y. Zhang, W. Chen, M. Zhang, and J. Zhu. Fast and accurate shift-reduce constituent parsing. In *ACL (1)*, pages 434–443, 2013. 3, 7

[34] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. *arXiv preprint arXiv:1511.03416*, 2015. 5, 7

[35] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2