

Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues

Bryan A. Plummer Arun Mallya Christopher M. Cervantes Julia Hockenmaier
Svetlana Lazebnik
University of Illinois at Urbana-Champaign

{bplumme2, amallya2, ccervan2, juliahmr, slazebni}@illinois.edu

Abstract

This paper presents a framework for localization or grounding of phrases in images using a large collection of linguistic and visual cues. We model the appearance, size, and position of entity bounding boxes, adjectives that contain attribute information, and spatial relationships between pairs of entities connected by verbs or prepositions. Special attention is given to relationships between people and clothing or body part mentions, as they are useful for distinguishing individuals. We automatically learn weights for combining these cues and at test time, perform joint inference over all phrases in a caption. The resulting system produces state of the art performance on phrase localization on the Flickr30k Entities dataset [33] and visual relationship detection on the Stanford VRD dataset [27].¹

1. Introduction

Today’s deep features can give reliable signals about a broad range of content in natural images, leading to advances in image-language tasks such as automatic captioning [6, 14, 16, 17, 42] and visual question answering [1, 8, 44]. A basic building block for such tasks is localization or grounding of individual phrases [6, 16, 17, 28, 33, 40, 42]. A number of datasets with phrase grounding information have been released, including Flickr30k Entities [33], ReferIt [18], Google Referring Expressions [29], and Visual Genome [21]. However, grounding remains challenging due to open-ended vocabularies, highly unbalanced training data, prevalence of hard-to-localize entities like clothing and body parts, as well as the subtlety and variety of linguistic cues that can be used for localization.

The goal of this paper is to accurately localize a bounding box for each entity (noun phrase) mentioned in a caption for a particular test image. We propose a joint localization objective for this task using a learned combination of single-phrase and phrase-pair cues. Evaluation is performed on the

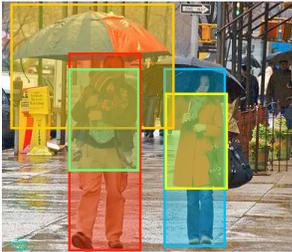
Input Sentence and Image	Cues	Examples
A man carries a baby under a red and blue umbrella next to a woman in a red jacket	1) Entities	man, baby, umbrella, woman, jacket
	2) Candidate Box Position	---
	3) Candidate Box Size	---
	4) Common Object Detectors	man → person baby → person woman → person
	5) Adjectives	umbrella → red umbrella → blue jacket → red
	6) Subject - Verb	(man, carries)
	7) Verb - Object	(carries, baby)
	8) Verbs	(man, carries, baby)
	9) Prepositions	(baby, under, umbrella) (man, next to, woman)
	10) Clothing & Body Parts	(woman, in, jacket)

Figure 1: Left: an image and caption, together with ground truth bounding boxes of entities (noun phrases). Right: a list of all the cues used by our system, with corresponding phrases from the sentence.

challenging recent Flickr30K Entities dataset [33], which provides ground truth bounding boxes for each entity in the five captions of the original Flickr30K dataset [43].

Figure 1 introduces the components of our system using an example image and caption. Given a noun phrase extracted from the caption, e.g., *red and blue umbrella*, we obtain single-phrase cue scores for each candidate box based on appearance (modeled with a phrase-region embedding as well as object detectors for common classes), size, position, and attributes (adjectives). If a pair of entities is connected by a verb (*man carries a baby*) or a preposition (*woman in a red jacket*), we also score the pair of corresponding candidate boxes using a spatial model. In addition, actions may modify the appearance of either the subject or the object (e.g., a man carrying a baby has a characteristic appearance, as does a baby being carried). To account for this, we learn subject-verb and verb-object appearance models for the constituent entities. We give special treatment to relationships between people, clothing, and body parts, as these are commonly used for describing individuals, and are also among the hardest entities for existing approaches to localize. To extract as complete a set of relationships as possible, we use natural language processing (NLP) tools to resolve pronoun references within a sentence: e.g., by analyzing the

¹Code: <https://github.com/BryanPlummer/pl-clc>

Method	Single Phrase Cues						Phrase-Pair Spatial Cues		Inference
	Phrase-Region Compatibility	Candidate Position	Candidate Size	Object Detectors	Adjectives	Verbs	Relative Position	Clothing & Body Parts	Joint Localization
Ours	✓	✓	✓	✓*	✓	✓	✓	✓	✓
(a) NonlinearSP [40]	✓	–	–	–	–	–	–	–	–
GroundedR [34]	✓	–	–	–	–	–	–	–	–
MCB [8]	✓	–	–	–	–	–	–	–	–
SCRC [12]	✓	✓	–	–	–	–	–	–	–
SMPL [41]	✓	–	–	–	–	–	✓*	–	✓
RtP [33]	✓	–	✓	✓*	✓*	–	–	–	–
(b) Scene Graph [15]	–	–	–	✓	✓	–	✓	–	✓
ReferIt [18]	–	✓	✓	✓	✓*	–	✓	–	–
Google RefExp [29]	✓	✓	✓	–	–	–	–	–	–

Table 1: Comparison of cues for phrase-to-region grounding. **(a)** Models applied to phrase localization on Flickr30K Entities. **(b)** Models on related tasks. * indicates that the cue is used in a limited fashion, i.e. [18, 33] restricted their adjective cues to colors, [41] only modeled possessive pronoun phrase-pair spatial cues ignoring verb and prepositional phrases, [33] and we limit the object detectors to 20 common categories.

sentence *A man puts his hand around a woman*, we can determine that the hand belongs to the man and introduce the respective pairwise term into our objective.

Table 1 compares the cues used in our work to those in other recent papers on phrase localization and related tasks like image retrieval and referring expression understanding. To date, other methods applied to the Flickr30K Entities dataset [8, 12, 34, 40, 41] have used a limited set of single-phrase cues. Information from the rest of the caption, like verbs and prepositions indicating spatial relationships, has been ignored. One exception is Wang *et al.* [41], who tried to relate multiple phrases to each other, but limited their relationships only to those indicated by possessive pronouns, not personal ones. By contrast, we use pronoun cues to the full extent by performing pronominal coreference. Also, ours is the only work in this area incorporating the visual aspect of verbs. Our formulation is most similar to that of [33], but with a larger set of cues, learned combination weights, and a global optimization method for simultaneously localizing all the phrases in a sentence.

In addition to our experiments on phrase localization, we also adapt our method to the recently introduced task of visual relationship detection (VRD) on the Stanford VRD dataset [27]. Given a test image, the goal of VRD is to detect all entities and relationships present and output them in the form (*subject, predicate, object*) with the corresponding bounding boxes. By contrast with phrase localization, where we are given a set of entities and relationships that are in the image, in VRD we do not know *a priori* which objects or relationships might be present. On this task, our model shows significant performance gains over prior work, with especially acute differences in zero-shot detection due to modeling cues with a vision-language embedding. This adaptability to never-before-seen examples is also a notable distinction between our approach and prior methods on related tasks (e.g. [7, 15, 18, 20]), which typically train their models on a set of predefined object categories, providing no support for out-of-vocabulary entities.

Section 2 discusses our global objective function for simultaneously localizing all phrases from the sentence and describes the procedure for learning combination weights. Section 3.1 details how we parse sentences to extract entities, relationships, and other relevant linguistic cues. Sections 3.2 and 3.3 define single-phrase and phrase-pair cost functions between linguistic and visual cues. Section 4 presents an in-depth evaluation of our cues on Flickr30K Entities [33]. Lastly, Section 5 presents the adaptation of our method to the VRD task [27].

2. Phrase localization approach

We follow the task definition used in [8, 12, 33, 34, 40, 41]: At test time, we are given an image and a caption with a set of entities (noun phrases), and we need to localize each entity with a bounding box. Section 2.1 describes our inference formulation, and Section 2.2 describes our procedure for learning the weights of different cues.

2.1. Joint phrase localization

For each image-language cue derived from a single phrase or a pair of phrases (Figure 1), we define a *cue-specific cost function* that measures its compatibility with an image region (small values indicate high compatibility). We will describe the cost functions in detail in Section 3; here, we give our test-time optimization framework for jointly localizing all phrases from a sentence.

Given a single phrase p from a test sentence, we score each region (bounding box) proposal b from the test image based on a linear combination of cue-specific cost functions $\phi_{\{1, \dots, K_S\}}(p, b)$ with learned weights w^S :

$$S(p, b; w^S) = \sum_{s=1}^{K_S} \mathbb{1}_s(p) \phi_s(p, b) w_s^S, \quad (1)$$

where $\mathbb{1}_s(p)$ is an indicator function for the availability of cue s for phrase p (e.g., an adjective cue would be available for the phrase *blue socks*, but would be unavailable for

socks by itself). As will be described in Section 3.2, we use 14 single-phrase cost functions: region-phrase compatibility score, phrase position, phrase size (one for each of the eight phrase types of [33]), object detector score, adjective, subject-verb, and verb-object scores.

For a pair of phrases with some relationship $r = (p, rel, p')$ and candidate regions b and b' , an analogous scoring function is given by a weighted combination of pairwise costs $\psi_{\{1, \dots, K_Q\}}(r, b, b')$:

$$Q(r, b, b'; w^Q) = \sum_{q=1}^{K_Q} \mathbb{1}_q(r) \psi_q(r, b, b') w_q^Q. \quad (2)$$

We use three pairwise cost functions corresponding to spatial classifiers for verb, preposition, and clothing and body parts relationships (Section 3.3).

We train all cue-specific cost functions on the training set and the combination weights on the validation set. At test time, given an image and a list of phrases $\{p_1, \dots, p_N\}$, we first retrieve top M candidate boxes for each phrase p_i using Eq. (1). Our goal is then to select one bounding box b_i out of the M candidates per each phrase p_i such that the following objective is minimized:

$$\min_{b_1, \dots, b_N} \left\{ \sum_{p_i} S(p_i, b_i) + \sum_{r_{ij}=(p_i, rel_{ij}, p_j)} Q(r_{ij}, b_i, b_j) \right\} \quad (3)$$

where phrases p_i and p_j (and respective boxes b_i and b_j) are related by some relationship rel_{ij} . This is a binary quadratic programming formulation inspired by [38]; we relax and solve it using a sequential QP solver in MATLAB. The solution gives a single bounding box hypothesis for each phrase. Performance is evaluated using Recall@1, or proportion of phrases where the selected box has Intersection-over-Union (IOU) ≥ 0.5 with the ground truth.

2.2. Learning scoring function weights

We learn the weights w^S and w^Q in Eqs. (1) and (2) by directly optimizing recall on the validation set. We start by finding the unary weights w^S that maximize the number of correctly localized phrases:

$$w^S = \arg \max_w \sum_{i=1}^N \mathbb{1}_{IOU \geq 0.5}(b_i^*, \hat{b}(p_i; w)), \quad (4)$$

where N is the number of phrases in the training set, $\mathbb{1}_{IOU \geq 0.5}$ is an indicator function returning 1 if the two boxes have IOU ≥ 0.5 , b_i^* is the ground truth bounding box for phrase p_i , $\hat{b}(p; w)$ returns the most likely box candidate for phrase p under the current weights, or, more formally, given a set of candidate boxes \mathcal{B} ,

$$\hat{b}(p; w) = \min_{b \in \mathcal{B}} S(p, b; w). \quad (5)$$

We optimize Eq. (4) using a derivative-free direct search method [22] (MATLAB's `fminsearch`). We randomly initialize the weights, keep the best weights after 20 runs based on validation set performance (takes just a few minutes to learn weights for all single phrase cues in our experiments).

Next, we fix w^S and learn the weights w^Q over phrase-pair cues in the validation set. To this end, we formulate an objective analogous to Eq. (4) for maximizing the number of correctly localized region pairs. Similar to Eq. (5), we define the function $\hat{\rho}(r; w)$ to return the best pair of boxes for the relationship $r = (p, rel, p')$:

$$\hat{\rho}(r; w) = \min_{b, b' \in \mathcal{B}} S(p, b; w^S) + S(p', b'; w^S) + Q(r, b, b'; w). \quad (6)$$

Then our pairwise objective function is

$$w^Q = \arg \max_w \sum_{k=1}^M \mathbb{I}_{PairIOU \geq 0.5}(\rho_k^*, \hat{\rho}(r_k; w)), \quad (7)$$

where M is the number of phrase pairs with a relationship, $\mathbb{I}_{PairIOU \geq 0.5}$ returns the number of correctly localized boxes (0, 1, or 2), and ρ_k^* is the ground truth box pair for the relationship $r_k = (p_k, rel_k, p'_k)$.

Note that we also attempted to learn the weights w^S and w^Q using standard approaches such as rank-SVM [13], but found our proposed direct search formulation to work better. In phrase localization, due to its Recall@1 evaluation criterion, only the correctness of one best-scoring candidate region for each phrase matters, unlike in typical detection scenarios, where one would like all positive examples to have better scores than all negative examples. The VRD task of Section 5 is a more conventional detection task, so there we found rank-SVM to be more appropriate.

3. Cues for phrase-region grounding

Section 3.1 describes how we extract linguistic cues from sentences. Sections 3.2 and 3.3 give our definitions of the two types of cost functions used in Eqs. (1) and (2): single phrase cues (SPC) measure the compatibility of a given phrase with a candidate bounding box, and phrase pair cues (PPC) ensure that pairs of related phrases are localized in a spatially coherent manner.

3.1. Extracting linguistic cues from captions

The Flickr30k Entities dataset provides annotations for Noun Phrase (NP) chunks corresponding to entities, but linguistic cues corresponding to adjectives, verbs, and prepositions must be extracted from the captions using NLP tools. Once these cues are extracted, they will be translated into visually relevant constraints for grounding. In particular, we will learn specialized detectors for adjectives, subject-verb, and verb-object relationships (Section 3.2). Also, because pairs of entities connected by a verb or preposition

have constrained layout, we will train classifiers to score pairs of boxes based on spatial information (Section 3.3).

Adjectives are part of NP chunks so identifying them is trivial. To extract other cues, such as verbs and prepositions that may indicate actions and spatial relationships, we obtain a constituent parse tree for each sentence using the Stanford parser [37]. Then, for possible relational phrases (prepositional and verb phrases), we use the method of Fidler *et al.* [7], where we start at the relational phrase and then traverse up the tree and to the left until we reach a noun phrase node, which will correspond to the first entity in an (*entity1*, *rel*, *entity2*) tuple. The second entity is given by the first noun phrase node on the right side of the relational phrase in the parse tree. For example, given the sentence *A boy running in a field with a dog*, the extracted NP chunks would be *a boy*, *a field*, *a dog*. The relational phrases would be (*a boy*, *running in*, *a field*) and (*a boy*, *with*, *a dog*).

Notice that a single relational phrase can give rise to multiple relationship cues. Thus, from (*a boy*, *running in*, *a field*), we extract the verb relation (*boy*, *running*, *field*) and prepositional relation (*boy*, *in*, *field*). An exception to this is a relational phrase where the first entity is a person and the second one is of the clothing or body part type,² e.g., (*a boy*, *running in*, *a jacket*). For this case, we create a single special pairwise relation (*boy*, *jacket*) that assumes that the second entity is attached to the first one and the exact relationship words do not matter, i.e., (*a boy*, *running in*, *a jacket*) and (*a boy*, *wearing*, *a jacket*) are considered to be the same. The attachment assumption can fail for phrases like (*a boy*, *looking at*, *a jacket*), but such cases are rare.

Finally, since pronouns in Flickr30k Entities are not annotated, we attempt to perform pronominal coreference (i.e., creating a link between a pronoun and the phrase it refers to) in order to extract a more complete set of cues. As an example, given the sentence *Ducks feed themselves*, initially we can only extract the subject-verb cue (*ducks*, *feed*), but we don’t know who or what they are feeding. Pronominal coreference resolution tells us that the ducks are themselves eating and not, say, feeding ducklings. We use a simple rule-based method similar to knowledge-poor methods [11, 31]. Given lists of pronouns by type,³ our rules attach each pronoun with at most one non-pronominal mention that occurs earlier in the sentence (an antecedent). We assume that subject and object pronouns often refer to the main subject (e.g. [*A dog*] *laying on the ground looks up at the dog standing over [him]*), reflexive and reciprocal pronouns refer to the nearest antecedent (e.g. [*A tennis player*] *readies [herself]*), and indefinite pronouns do not refer to a previously described entity. It must be noted that

²Each NP chunk from the Flickr30K dataset is classified into one of eight phrase types based on the dictionaries of [33].

³Relevant pronoun types are subject, object, reflexive, reciprocal, relative, and indefinite.

compared with verb and prepositional relationships, relatively few additional cues are extracted using this procedure (432 pronoun relationships in the test set and 13,163 in the train set, while the counts for the other relationships are on the order of 10K and 300K).

3.2. Single Phrase Cues (SPCs)

Region-phrase compatibility: This is the most basic cue relating phrases to image regions based on appearance. It is applied to every test phrase (i.e., its indicator function in Eq. (1) is always 1). Given phrase p and region b , the cost $\phi_{CCA}(p, b)$ is given by the cosine distance between p and b in a joint embedding space learned using normalized Canonical Correlation Analysis (CCA) [10]. We use the same procedure as [33]. Regions are represented by the fc7 activations of a Fast-RCNN model [9] fine-tuned using the union of the PASCAL 2007 and 2012 trainval sets [5]. After removing stopwords, phrases are represented by the HGLMM fisher vector encoding [19] of word2vec [30].

Candidate position: The location of a bounding box in an image has been shown to be predictive of the kinds of phrases it may refer to [4, 12, 18, 23]. We learn location models for each of the eight broad phrase types specified in [33]: people, clothing, body parts, vehicles, animals, scenes, and a catch-all “other.” We represent a bounding box by its centroid normalized by the image size, the percentage of the image covered by the box, and its aspect ratio, resulting in a 4-dim. feature vector. We then train a support vector machine (SVM) with a radial basis function (RBF) kernel using LIBSVM [2]. We randomly sample EdgeBox [46] proposals with IOU < 0.5 with the ground truth boxes for negative examples. Our scoring function is

$$\phi_{pos}(p, b) = -\log(\text{SVM}_{type(p)}(b)),$$

where $\text{SVM}_{type(p)}$ returns the probability that box b is of the phrase type $type(p)$ (we use Platt scaling [32] to convert the SVM output to a probability).

Candidate size: People have a bias towards describing larger, more salient objects, leading prior work to consider the size of a candidate box in their models [7, 18, 33]. We follow the procedure of [33], so that given a box b with dimensions normalized by the image size, we have

$$\phi_{size_{type(p)}}(p, b) = 1 - b_{width} \times b_{height}.$$

Unlike phrase position, this cost function does not use a trained SVM per phrase type. Instead, each phrase type is its own feature and the corresponding indicator function returns 1 if that phrase belongs to the associated type.

Detectors: CCA embeddings are limited in their ability to localize objects because they must account for a wide range of phrases and because they do not use negative examples

during training. To compensate for this, we use Fast R-CNN [9] to learn three networks for common object categories, attributes, and actions. Once a detector is trained, its score for a region proposal b is

$$\phi_{det}(p, b) = -\log(\text{softmax}_{det}(p, b)),$$

where $\text{softmax}_{det}(p, b)$ returns the output of the softmax layer for the object class corresponding to p . We manually create dictionaries to map phrases to detector categories (e.g., man, woman, *etc.* map to ‘person’), and the indicator function for each detector returns 1 only if one of the words in the phrase exists in its dictionary. If multiple detectors for a single cue type are appropriate for a phrase (e.g., *a black and white shirt* would have two adjective detectors fire, one for each color), the scores are averaged. Below, we describe the three detector networks used in our model. Complete dictionaries can be found in supplementary material.

Objects: We use the dictionary of [33] to map nouns to the 20 PASCAL object categories [5] and fine-tune the network on the union of the PASCAL VOC 2007 and 2012 trainval sets. At test time, when we run a detector for a phrase that maps to one of these object categories, we also use bounding box regression to refine the original region proposals. Regression is not used for the other networks below.

Adjectives: Adjectives found in phrases, especially color, provide valuable attribute information for localization [7, 15, 18, 33]. The Flickr30K Entities baseline approach [33] used a network trained for 11 colors. As a generalization of that, we create a list of adjectives that occur at least 100 times in the training set of Flickr30k. After grouping together similar words and filtering out non-visual terms (e.g., *adventurous*), we are left with a dictionary of 83 adjectives. As in [33], we consider color terms describing people (*black man, white girl*) to be separate categories.

Subject-Verb and Verb-Object: Verbs can modify the appearance of both the subject and the object in a relation. For example, knowing that a person is riding a horse can give us better appearance models for finding both the person and the horse [35, 36]. As we did with adjectives, we collect verbs that occur at least 100 times in the training set, group together similar words, and filter out those that don’t have a clear visual aspect, resulting in a dictionary of 58 verbs. Since a person running looks different than a dog running, we subdivide our verb categories by phrase type of the subject (resp. object) if that phrase type occurs with the verb at least 30 times in the train set. For example, if there are enough animal-running occurrences, we create a new category with instances of all animals running. For the remaining phrases, we train a catch-all detector over all the phrases related to that verb. Following [35], we train separate detectors for subject-verb and verb-object relationships, resulting in dictionary sizes of 191 (resp. 225). We also attempted to

learn subject-verb-object detectors as in [35, 36], but did not see a further improvement.

3.3. Phrase-Pair Cues (PPCs)

So far, we have discussed cues pertaining to a single phrase, but relationships between pairs of phrases can also provide cues about their relative position. We denote such relationships as tuples $(p_{left}, rel, p_{right})$ with *left, right* indicating on which side of the relationship the phrases occur. As discussed in Section 3.1, we consider three distinct types of relationships: verbs (*man, riding, horse*), prepositions (*man, on, horse*), and clothing and body parts (*man, wearing, hat*). For each of the three relationship types, we group phrases referring to people but treat all other phrases as distinct, and then gather all relationships that occur at least 30 times in the training set. Then we learn a spatial relationship model as follows. Given a pair of boxes with coordinates $b = (x, y, w, h)$ and $b' = (x', y', w', h')$, we compute a four-dim. feature

$$[(x - x')/w, (y - y')/h, w'/w, h'/h], \quad (8)$$

and concatenate it with combined SPC scores $S(p_{left}, b)$, $S(p_{right}, b')$ from Eq. (1). To obtain negative examples, we randomly sample from other box pairings with IOU < 0.5 with the ground truth regions from that image. We train an RBF SVM classifier with Platt scaling [32] to obtain a probability output. This is similar to the method of [15], but rather than learning a Gaussian Mixture Model using only positive data, we learn a more discriminative model. Below are details on the three types of relationship classifiers.

Verbs: Starting with our dictionary of 58 verb detectors and following the above procedure of identifying all relationships that occur at least 30 times in the training set, we end up with 260 $(p_{left}, rel_{verb}, p_{right})$ SVM classifiers.

Prepositions: We first gather a list of prepositions that occur at least 100 times in the training set, combine similar words, and filter out words that do not indicate a clear spatial relationship. This yields eight prepositions (*in, on, under, behind, across, between, onto, and near*) and 216 $(p_{left}, rel_{prep}, p_{right})$ relationships.

Clothing and body part attachment: We collect $(p_{left}, rel_{c\&bp}, p_{right})$ relationships where the left phrase is always a person and the right phrase is from the clothing or body part type and learn 207 such classifiers. As discussed in Section 3.1, this relationship type takes precedence over any verb or preposition relationships that may also hold between the same phrases.

4. Experiments on Flickr30k Entities

4.1. Implementation details

We utilize the provided train/test/val split of 29,873 training, 1,000 validation, and 1,000 testing images [33].

Method	Accuracy
(a) Single-phrase cues	
CCA	43.09
CCA+Det	45.29
CCA+Det+Size	51.45
CCA+Det+Size+Adj	52.63
CCA+Det+Size+Adj+Verbs	54.51
CCA+Det+Size+Adj+Verbs+Pos (SPC)	55.49
(b) Phrase pair cues	
SPC+Verbs	55.53
SPC+Verbs+Preps	55.62
SPC+Verbs+Preps+C&BP (SPC+PPC)	55.85
(c) State of the art	
SMPL [41]	42.08
NonlinearSP [40]	43.89
Grounder [34]	47.81
MCB [8]	48.69
RtP [33]	50.89

Table 2: Phrase-region grounding performance on the Flickr30k Entities dataset. (a) Performance of our single-phrase cues (Sec. 3.2). (b) Further improvements by adding our pairwise cues (Sec. 3.3). (c) Accuracies of competing state-of-the-art methods. This comparison excludes concurrent work that was published after our initial submission [3].

Following [33], our region proposals are given by the top 200 EdgeBox [46] proposals per image. At test time, given a sentence and an image, we first use Eq. (1) to find the top 30 candidate regions for each phrase after performing non-maximum suppression using a 0.8 IOU threshold. Restricted to these candidates, we optimize Eq. (2) to find a globally consistent mapping of phrases to regions.

Consistent with [33], we only evaluate localization for phrases with a ground truth bounding box. If multiple bounding boxes are associated with a phrase (e.g., four individual boxes for *four men*), we represent the phrase as the union of its boxes. For each image and phrase in the test set, the predicted box must have at least 0.5 IOU with its ground truth box to be deemed successfully localized. As only a single candidate is selected for each phrase, we report the proportion of correctly localized phrases (i.e. Recall@1).

4.2. Results

Table 2 reports our overall localization accuracy for combinations of cues and compares our performance to the state of the art. Object detectors, reported on the second line of Table 2(a), show a 2% overall gain over the CCA baseline. This includes the gain from the detector score as well as the bounding box regressor trained with the detector in the Fast R-CNN framework [9]. Adding adjective, verb, and size cues improves accuracy by a further 9%. Our last cue in Table 2(a), position, provides an additional 1% improvement.

We can see from Table 2(b) that the spatial cues give only a small overall boost in accuracy on the test set, but that is due to the relatively small number of phrases to which they apply. In Table 4 we will show that the localization improvement on the affected phrases is much larger.

Table 2(c) compares our performance to the state of the art. The method most similar to ours is our earlier

model [33], which we call RtP here. RtP relies on a subset of our single-phrase cues (region-phrase CCA, size, object detectors, and color adjectives), and localizes each phrase separately. The closest version of our current model to RtP is CCA+Det+Size+Adj, which replaces the 11 colors of [33] with our more general model for 83 adjectives, and obtains almost 2% better performance. Our full model is 5% better than RtP. It is also worth noting that a rank-SVM model [13] for learning cue combination weights gave us 8% worse performance than the direct search scheme of Section 2.2.

Table 3 breaks down the comparison by phrase type. Our model has the highest accuracy on most phrase types, with scenes being the most notable exception, for which Grounder [34] does better. However, Grounder uses Selective Search proposals [39], which have an upper bound performance that is 7% higher on scene phrases despite using half as many proposals. Although body parts have the lowest localization accuracy at 25.24%, this represents an 8% improvement in accuracy over prior methods. However, only around 62% of body part phrases have a box with high enough IOU with the ground truth, showing a major area of weakness of category-independent proposal methods. Indeed, if we were to augment our EdgeBox region proposals with ground truth boxes, we would get an overall improvement in accuracy of about 9% for the full system.

Since many of the cues apply to a small subset of the phrases, Table 4 details the performance of cues over only the phrases they affect. As a baseline, we compare against the combination of cues available for all phrases: region-phrase CCA, position, and size. To have a consistent set of regions, the baseline also uses improved boxes from bounding box regressors trained along with the object detectors. As a result, the object detectors provide less than 2% gain over the baseline for the phrases on which they are used, suggesting that the regression provides the majority of the gain from CCA to CCA+Det in Table 2. This also confirms that there is significant room for improvement in selecting candidate regions. By contrast, adjective, subject-verb, and verb-object detectors show significant gains, improving over the baseline by 6-7%.

The right side of Table 4 shows the improvement on phrases due to phrase pair cues. Here, we separate the phrases that occur on the left side of the relationship, which corresponds to the subject, from the phrases on the right side. Our results show that the subject, is generally easier to localize. On the other hand, clothing and body parts show up mainly on the right side of relationships and they tend to be small. It is also less likely that such phrases will have good candidate boxes – recall from Table 3 that body parts have a performance upper bound of only 62%. Although they affect relatively few test phrases, all three of our relationship classifiers show consistent gains over the SPC

	People	Clothing	Body Parts	Animals	Vehicles	Instruments	Scene	Other
#Test	5,656	2,306	523	518	400	162	1,619	3,374
SMPL [41]	57.89	34.61	15.87	55.98	52.25	23.46	34.22	26.23
GroundeR [34]	61.00	38.12	10.33	62.55	68.75	36.42	58.18	29.08
RtP [33]	64.73	46.88	17.21	65.83	68.75	37.65	51.39	31.77
SPC+PPC (ours)	71.69	50.95	25.24	76.25	66.50	35.80	51.51	35.98
Upper Bound	97.72	83.13	61.57	91.89	94.00	82.10	84.37	81.06

Table 3: Comparison of phrase localization performance over phrase types. Upper Bound refers to the proportion of phrases of each type for which there exists a region proposal having at least 0.5 IOU with the ground truth.

Method	Single Phrase Cues (SPC)				Phrase-Pair Cues (PPC)					
	Object Detectors	Adjectives	Subject-Verb	Verb-Object	Verbs		Prepositions		Clothing & Body Parts	
					Left	Right	Left	Right	Left	Right
Baseline	74.25	57.71	69.68	40.70	78.32	51.05	68.97	55.01	81.01	50.72
+Cue	75.78	64.35	75.53	47.62	78.94	51.33	69.74	56.14	82.86	52.23
#Test	4,059	3,809	3,094	2,398	867	858	780	778	1,464	1,591
#Train	114,748	110,415	94,353	71,336	26,254	25,898	23,973	23,903	42,084	45,496

Table 4: Breakdown of performance for individual cues restricted only to test phrases to which they apply. For SPC, Baseline is given by CCA+Position+Size. For PPC, Baseline is the full SPC model. For all comparisons, we use the improved boxes from bounding box regression on top of object detector output. PPC evaluation is split by which side of the relationship the phrases occur on. The bottom two rows show the numbers of affected phrases in the test and training sets. For reference, there are 14.5k visual phrases in the test set and 427k visual phrases in the train set.

model. This is encouraging given that many of the relationships that are used on the validation set to learn our model parameters do not occur in the test set (and vice versa).

Figure 2 provides a qualitative comparison of our output with the RtP model [33]. In the first example, the prediction for the dog is improved due to the subject-verb classifier for *dog jumping*. For the second example, pronominal coreference resolution (Section 3.1) links *each other* to *two men*, telling us that not only is a man hitting something, but also that another man is being hit. In the third example, the RtP model is not able to locate the woman’s blue stripes in her hair despite having a model for *blue*. Our adjective detectors take into account *stripes* as well as *blue*, allowing us to correctly localize the phrase, even though we still fail to localize the hair. Since the blue stripes and hair should co-locate, a method for obtaining co-referent entities would further improve performance on such cases. In the last example, the RtP model makes the same incorrect prediction for the two men. However, our spatial relationship between the first man and his gray sweater helps us correctly localize him. We also improve our prediction for the shopping cart.

5. Visual Relationship Detection

In this section, we adapt our framework to the recently introduced Visual Relationship Detection (VRD) benchmark of Lu *et al.* [27]. Given a test image without any text annotations, the task of VRD is to detect all entities and relationships present and output them in the form (*subject, predicate, object*) with the corresponding bounding boxes. A relationship detection is judged to be correct if it exists in the image and both the subject and object boxes have $\text{IOU} \geq 0.5$ with their respective ground truth. In contrast to phrase grounding, where we are given a set of entities and relationships that are assumed to be in the image, here we

do not know *a priori* which objects or relationships might be present. On the other hand, the VRD dataset is easier than Flickr30K Entities in that it has a limited vocabulary of 100 object classes and 70 predicates annotated in 4000 training and 1000 test images.

Given the small fixed class vocabulary, it would seem advantageous to train 100 object detectors on this dataset, as was done by Lu *et al.* [27]. However, the training set is relatively small, the class distribution is unbalanced, and there is no validation set. Thus, we found that training detectors and then relationship models on the same images causes overfitting because the detector scores on the training images are overconfident. We obtain better results by training all appearance models using CCA, which also takes into account semantic similarity between category names and is trivially extendable to previously unseen categories. Here, we use fc7 features from a Fast RCNN model trained on MSCOCO [26] due to the larger range of categories than PASCAL, and word2vec for object and predicate class names. We train the following CCA models:

1. CCA(entity box, entity class name): this is the equivalent to region-phrase CCA in Section 3.2 and is used to score both candidate subject and object boxes.
2. CCA(subject box, [subject class name, predicate class name]): analogous to subject-verb classifiers of Section 3.2. The 300-dimensional word2vec features of subject and predicate class names are concatenated.
3. CCA(object box, [predicate class name, object class name]): analogous to verb-object classifiers of Section 3.2.
4. CCA(union box, predicate class name): this model measures the compatibility between the bounding box of both subject and object and the predicate name.
5. CCA(union box, [subject class name, predicate class name, object class name]).

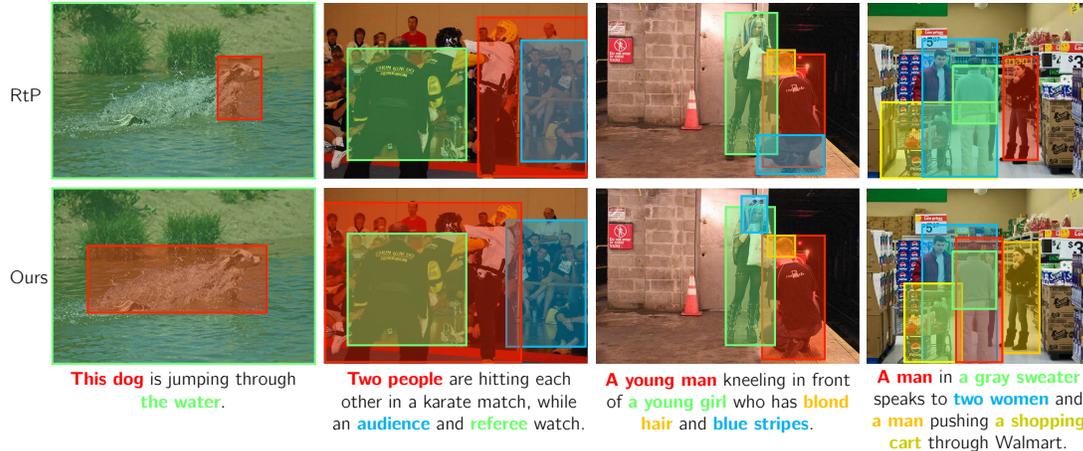


Figure 2: Example results on Flickr30k Entities comparing our SPC+PPC model’s output with the RtP model [33]. See text for discussion.

Note that models 4 and 5 had no analogue in our phrase localization system. On that task, entities were known to be in the image and relationships simply provided constraints, while here we need to predict which relationships exist. To make predictions for predicates and relationships (which is the goal of models 4 and 5), it helps to see both the subject and object regions. Union box features were also less useful for phrase localization due to the larger vocabularies and relative scarcity of relationships in that task.

Each candidate relationship gets six CCA scores (model 1 above is applied both to the subject and the object). In addition, we compute size and position scores as in Section 3.2 for subject and object, and a score for a pairwise spatial SVM trained to predict the predicate based on the four-dimensional feature of Eq. (8). This yields an 11-dim. feature vector. By contrast with phrase localization, our features for VRD are dense (always available for every relationship).

In Section 2.2 we found feature weights by maximizing our recall metric. Here we have a more conventional detection task, so we obtain better performance by training a linear rank-SVM model [13] to enforce that correctly detected relationships are ranked higher than negative detections (where either box has < 0.5 IOU with the ground truth). We use the test set object detections (just the boxes, not the scores) provided by [27] to directly compare performance with the same candidate regions. During testing, we produce a score for every ordered pair of detected boxes and all possible predicates, and retain the top 10 predicted relationships per pair of (subject, object) boxes.

Consistent with [27], Table 5 reports recall, $R@ \{100, 50\}$, or the portion of correctly localized relationships in the top 100 (resp. 50) ranked relationships in the image. The right side shows performance for relationships that have not been encountered in the training set. Our method clearly outperforms that of Lu *et al.* [27], which uses separate visual, language, and relationship likelihood cues. We also

Method	Rel. Det.		Zero-shot Rel. Det.	
	R@100	R@50	R@100	R@50
(a) Visual Only Model [27]	1.85	1.58	0.78	0.67
Visual + Language + Likelihood Model [27]	14.70	13.86	3.52	3.13
VTransE [45]	15.20	14.07	2.14	1.71
(b) CCA	13.69	10.08	11.12	6.59
CCA + Size	14.05	10.36	11.46	6.76
CCA + Size + Position	18.37	15.08	13.43	9.67

Table 5: Relationship detection recall at different thresholds ($R@ \{100, 50\}$). CCA refers to the combination of six CCA models (see text). Position refers to the combination of individual box position and pairwise spatial classifiers. This comparison excludes concurrent work that was published after our initial submission [24, 25].

outperform Zhang *et al.* [45], which combines object detectors, visual appearance, and object position in a single neural network. We observe that cues based on object class and relative subject-object position provide a noticeable boost in performance. Further, due to using CCA with multi-modal embeddings, we generalize better to unseen relationships.

6. Conclusion

This paper introduced a framework incorporating a comprehensive collection of image- and language-based cues for visual grounding and demonstrated significant gains over the state of the art on two tasks: phrase localization on Flickr30k Entities and relationship detection on the VRD dataset. For the latter task, we got particularly pronounced gains for the zero-shot learning scenario. In future work, we would like to train a single network for combining multiple cues. Doing this in a unified end-to-end fashion is challenging, since one needs to find the right balance between parameter sharing and specialization or fine-tuning required by individual cues. To this end, our work provides a strong baseline and can help to inform future approaches.

Acknowledgments. This work was partially supported by NSF grants 1053856, 1205627, 1405883, 1302438, and 1563727, Xerox UAC, the Sloan Foundation, and a Google Research Award.

References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 1
- [2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. 4
- [3] K. Chen, R. Kovvuri, J. Gao, and R. Nevatia. MSRC: Multimodal spatial regression with semantic context for phrase grounding. In *ICMR*, 2017. 6
- [4] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Heber. An empirical study of context in object detection. In *CVPR*, 2009. 4
- [5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>, 2012. 4, 5
- [6] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015. 1
- [7] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In *CVPR*, 2013. 2, 4, 5
- [8] 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 *EMNLP*, 2016. 1, 2, 6
- [9] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 4, 5, 6
- [10] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014. 4
- [11] S. Harabagiu and S. Maiorano. Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In *Proceedings of the ACL-99 Workshop on the relation of discourse/dialogue structure and reference*, pages 29–38, 1999. 4
- [12] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *CVPR*, 2016. 2, 4
- [13] T. Joachims. Training linear svms in linear time. In *SIGKDD*, 2006. 3, 6, 8
- [14] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 1
- [15] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 2, 5
- [16] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1
- [17] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 1
- [18] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1, 2, 4, 5
- [19] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vector. In *CVPR*, 2015. 4
- [20] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 2
- [21] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1
- [22] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112147, 1998. 3
- [23] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei. Object bank: An object-level image representation for high-level visual recognition. *IJCV*, 107(1):20–39, 2014. 4
- [24] Y. Li, W. Ouyang, X. Wang, and X. Tang. ViP-CNN: Visual phrase guided convolutional neural network. In *CVPR*, 2017. 8
- [25] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, 2017. 8
- [26] 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. 7
- [27] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 1, 2, 7, 8
- [28] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015. 1
- [29] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1, 2
- [30] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013. 4
- [31] R. Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 869–875. Association for Computational Linguistics, 1998. 4
- [32] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999. 4, 5
- [33] 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. *IJCV*, 123(1):74–93, 2017. 1, 2, 3, 4, 5, 6, 7, 8

- [34] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 2, 6, 7
- [35] F. Sadeghi, S. K. Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, 2015. 5
- [36] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 5
- [37] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng. Parsing With Compositional Vector Grammars. In *ACL*, 2013. 4
- [38] J. Tighe, M. Niethammer, and S. Lazebnik. Scene parsing with object instances and occlusion ordering. In *CVPR*, 2014. 3
- [39] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 104(2), 2013. 6
- [40] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 1, 2, 6
- [41] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng. Structured matching for phrase localization. In *ECCV*, 2016. 2, 6, 7
- [42] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1
- [43] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 1
- [44] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual Madlibs: Fill in the blank Image Generation and Question Answering. In *ICCV*, 2015. 1
- [45] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017. 8
- [46] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 4, 6