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. 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 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
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 prononimal 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 \( \phi_{s} \) with learned weights \( w_{s} \):

\[
S(p, b; w^S) = \sum_{s=1}^{K_S} \mathbb{1}_{s}(p) \phi_{s}(p, b)w_{s}^{S},
\]

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

| 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. | Single Phrase Cues | Phrase-Pair Spatial Cues | Inference |
|---|---|---|---|
| Method | Phrase-Region Compatibility | Candidate Position | Candidate Size | Object Detectors | Adjectives | Verbs | Relative Position | Clothing & Body Parts | Joint Localization |
| Ours | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (a) NonlinearSP [40] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| GroundR [34] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| MCB [8] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| SCRC [12] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| SMPL [41] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| RiP [33] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (b) Scene Graph [15] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ReferIt [18] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Google RefExp [29] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
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 $\psi_{q=1}^{K_q}(r, b, b')$:

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

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 $\{p_1, \cdots, p_N\}$, 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:

$$\min_{b_1, \cdots, b_N} \left\{ \sum_{p_i} S(p_i, b_i) + \sum_{r_{ij} = \text{rel}_{ij} \in p_i} Q(r_{ij}, b_i, b_j) \right\} \quad (3)$$

where phrases $p_i$ and $p_j$ (and respective boxes $b_i$ and $b_j$) are related by some relationship $\text{rel}_{ij}$. 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) $\geq 0.5$ 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:

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

where $N$ is the number of phrases in the training set, $\mathbb{I}_{\text{IOU} \geq 0.5}$ is an indicator function returning 1 if the two boxes have IOU $\geq 0.5$, $b^*_i$ 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 $B$,

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

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 sets. To this end, we formulate an objective analogous to Eq. (4) for maximizing the number of correctly localized region sets.

Then our pairwise objective function is

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

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

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, e.g., (a boy, wearing, 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, 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] reads [herself]), and indefinite pronouns do not refer to a previously described entity. It must be noted that 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 HGLLMM 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

\[
\phi_{pos}(p, b) = -\log(SVM_{type(p)}(b)),
\]

where 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.

---

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.
Adjectives: Adjectives found in phrases, especially color, regression is not used for the other networks below. We use the dictionary of [33] to map nouns to 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 Appendix B.

**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

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

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. Complete dictionaries can be found in Appendix C.

**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 \((p_{left}, rel_{c,b}, p_{right})\) 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.
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].

| 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 |
| RnP [33] | 50.89 |

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]. 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 RnP here. RnP 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 RnP 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 RnP. 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
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) |
|------------------|--------------------------|------------------------|
|                  | #Train                   | #Test                  | Left | Right | Left | Right | Left | Right | Left | Right | Left | Right |
| Baseline         | 114,748                  | 4,059                  | 57.71 | 57.71 | 69.68 | 40.70 | 878.32 | 51.05 | 51.33 | 51.33 | 51.33 | 81.01 | 50.72 |
| +Cue             | 110,415                  | 3,809                  | 64.35 | 74.75 | 51.33 | 47.62 | 878.94 | 51.33 | 69.79 | 56.14 | 82.86 | 52.23 |

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, pronoun 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.

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 set and training sets. For reference, there are 14.5k visual phrases in the test set and 427k visual phrases in the train set.

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 IOU ≥ 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 is relatively small, the class distribution is unbalanced, and there is no validation set. 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]...
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 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. Qualitative examples and associated discussion can be found in Appendix A.

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.

[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.

[3] K. Chen, R. Kovvuri, J. Gao, and R. Nevatia. Msric: Multimodal spatial regression with semantic context for phrase grounding. In ICME, 2017.

[4] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In CVPR, 2009.

[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.

[6] H. Fang, S. Gupta, F. Iandola, R. Sreevastava, 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.

[7] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In CVPR, 2013.

[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.

[9] R. Girshick. Fast r-cnn. In ICCV, 2015.

[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.

[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.

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

[13] T. Joachims. Training linear svms in linear time. In SIGKDD, 2006.

[14] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In CVPR, 2016.

[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.

[16] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.

[17] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In NIPS, 2014.

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

[19] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vector. In CVPR, 2015.

[20] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In CVPR, 2014.

[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 through crowdsourced image annotations. IJCV, 2017.

[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):112–124, 1998.

[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.

[24] Y. Li, W. Ouyang, X. Wang, and X. Tang. ViP-CNN: Visual phrase guided convolutional neural network. In CVPR, 2017.

[25] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In CVPR, 2017.
[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, 9
[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. Mikhalcea, 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, 9
[46] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In ECCV, 2014. 4, 6
A. Visualization of detected relationships (VRD Dataset)

Below are some example detections on the VRD test set. Figure 3 shows some of the highly confident and correctly localized detections. We detect different types of relationships - spatial (post, behind, car), (sky, above, laptop), (laptop, on, table), clothing (person, wear, hat), (person, has, shorts), and actions (person, ride, skateboard).

![Images of detected relationships](image)

**Figure 3:** Highly confident and correctly localized relationships on the VRD dataset.

Figure 4 shows detections which were marked as negatives by the evaluation code as these relationships were not annotated in the corresponding images. However, note that these predictions are logically correct. The mouse is indeed next to the laptop (leftmost, first row), and the laptop is under the sky (middle, first row). Further, in the leftmost, second row image of Figure 3, the relationship (person, has, shorts) was marked as present, whereas the middle, second row image in Figure 4 has (person, has, hat) marked as absent, which indicates a lapse in annotation.

Figure 5 shows examples of wrongly detected relationships. Some of these relationships are logically implausible such as (hat, hold, surfboard) (leftmost, first row), while others such as (jeans, on, table) (middle, first row), while plausible, aren’t contextually true in the image. Other failure modes include incorrect detections such as the sky in the (rightmost, first row) image and the phone in the (leftmost, second row) image.
Figure 4: Plausible and logically correct detected relationships, penalized as negatives due to lack of annotations in the VRD dataset.

Figure 5: Falsely detected relationships on the VRD dataset. Mistakes are either due to incorrect localization of objects, prediction of implausible relationships, contextually incorrect relationships, or a combination of mistakes.
### B. List of detector classes from Flickr30k Entities

#### B.1. Adjectives

| 1) white      | 2) people-white | 3) female | 4) empty | 5) new | 6) black | 7) people-black |
|---------------|-----------------|-----------|----------|--------|----------|-----------------|
| 8) grassy     | 9) wet          | 10) colored | 11) red  | 12) people-red | 13) sunny | 14) smiling    |
| 15) professional | 16) brown      | 17) people-blond | 18) snowy | 19) african | 20) indoor | 21) gray       |
| 22) people-blue | 23) male       | 24) indian  | 25) oriental | 26) blond | 27) people-green | 28) crowded   |
| 29) bald       | 30) cold        | 31) blue    | 32) people-yellow | 33) shirtless | 34) american | 35) hot         |
| 36) green      | 37) young       | 38) dirt    | 39) dark-haired | 40) dark-skinned | 41) orange | 42) younger     |
| 43) paved      | 44) teenage     | 45) cloudy  | 46) pink  | 47) older | 48) rocky | 49) urban       |
| 50) military   | 51) purple      | 52) asian   | 53) hard  | 54) light | 55) hooded | 56) yellow      |
| 57) dark       | 58) beautiful   | 59) sandy   | 60) adult | 61) golden | 62) elderly | 63) bright      |
| 64) chinese    | 65) little      | 66) tan     | 67) old   | 68) concrete | 69) outdoors | 70) long        |
| 71) colorful   | 72) wooden      | 73) full    | 74) plastic | 75) tall  | 76) striped | 77) middle-aged  |
| 78) multicolored | 79) bearded    | 80) huge    | 81) short | 82) high  | 83) top    |                 |

#### B.2. Subject-Verb

| 1) animals-catching | 2) animals-climbing | 3) animals-digging | 4) animals-fighting | 5) animals-flying |
|---------------------|---------------------|--------------------|--------------------|-------------------|
| 6) animals-holding  | 7) animals-jumping  | 8) animals-playing | 9) animals-running | 10) animals-sitting |
| 11) animals-sleeping| 12) animals-splashing| 13) animals-standing| 14) animals-swimming| 15) animals-walking |
| 16) bodyparts-holding| 17) bodyparts-sitting| 18) bodyparts-walking| 19) clothing-climbing| 20) clothing-dancing |
| 21) clothing-eating | 22) clothing-holding| 23) clothing-jumping| 24) clothing-performing| 25) clothing-playing |
| 26) clothing-posing | 27) clothing-reading| 28) clothing-riding| 29) clothing-running| 30) clothing-singing |
| 31) clothing-sitting| 32) clothing-sleeping| 33) clothing-smiling| 34) clothing-standing| 35) clothing-talking |
| 36) clothing-walking| 37) clothing-working| 38) instruments-singing| 39) other-cooking| 40) other-drinking |
| 41) other-eating    | 42) other-flying    | 43) other-holding  | 44) other-jumping  | 45) other-performing |
| 46) other-playing   | 47) other-pointing  | 48) other-posing   | 49) other-reading  | 50) other-riding   |
| 51) other-running   | 52) other-singing   | 53) other-sitting  | 54) other-sleeping | 55) other-smiling  |
| 56) other-standing  | 57) other-talking   | 58) other-throwing | 59) other-walking  | 60) other-working  |
| 61) other-writing   | 62) people-blowing  | 63) people-catching| 64) people-cleaning| 65) people-climbing |
| 66) people-cooking  | 67) people-cutting  | 68) people-dancing | 69) people-digging | 70) people-drawing |
| 71) people-drinking | 72) people-driving  | 73) people-eating  | 74) people-falling | 75) people-fighting |
| 76) people-fishing  | 77) people-flying   | 78) people-hiking  | 79) people-hit     | 80) people-holding |
| 81) people-hugging  | 82) people-juggling | 83) people-jumping | 84) people-kicking | 85) people-kissing |
| 86) people-kneeling | 87) people-laughing | 88) people-painting| 89) people-performing| 90) people-playing |
| 91) people-pointing | 92) people-posing  | 93) people-pushing | 94) people-reaching| 95) people-reading |
| 96) people-riding   | 97) people-running | 98) people-serving | 99) people-shopping| 100) people-singing |
| 101) people-sitting | 102) people-skimming| 103) people-sleeping| 104) people-sliding| 105) people-smiling |
| 106) people-smoking | 107) people-splashing| 108) people-standing| 109) people-surfing| 110) people-sweeping |
| 111) people-swimming| 112) people-swinging| 113) people-talking| 114) people-throwing| 115) people-touches |
| 116) people-walking | 117) people-waving  | 118) people-working| 119) people-writing| 120) scene-eating  |
| 121) scene-holding  | 122) scene-playing | 123) scene-reading | 124) scene-running | 125) scene-sitting  |
| 126) scene-standing | 127) scene-talking | 128) scene-walking | 129) vehicles-driving| 130) vehicles-holding |
| 131) vehicles-running| 132) vehicles-sitting| 133) vehicles-throwing| 134) sitting     | 135) holding      |
| 136) playing       | 137) standing      | 138) walking       | 139) running       | 140) riding       |
| 141) jumping       | 142) working       | 143) talking       | 144) performing    | 145) eating       |
| 146) posing        | 147) climbing      | 148) hiking        | 149) reading       | 150) dancing      |
| 151) smiling       | 152) singing       | 153) sleeping      | 154) pushing       | 155) swimming     |
| 156) throwing      | 157) painting      | 158) driving       | 159) cooking       | 160) cutting      |
| 161) cleaning      | 162) serving       | 163) swinging      | 164) laughing      | 165) kicking      |
| 166) hit           | 167) fighting      | 168) juggling      | 167) drinking      | 170) kissing      |
| 171) pointing      | 172) blowing       | 173) sliding       | 174) drinking      | 175) fishing      |
| 176) writing       | 177) skiing        | 178) catching      | 179) kneelingle     | 180) hugging      |
| 181) digging       | 182) smoking       | 183) shopping      | 184) surfing       | 185) waving       |
| 186) sweeping      | 187) falling       | 188) reaching      | 189) drawing       | 190) splashing    |
| 191) touches       |                    |                    |                    |                   |
B.3. Verb-Object

1) other-blowing 2) other-catching 3) scene-catching 4) other-cleaning
5) scene-climbing 6) bodyparts-climbing 7) other-climbing 8) scene-climbing
9) other-cooking 10) bodyparts-cooking 11) bodyparts-cutting 12) other-cutting
13) clothing-dancing 14) other-dancing 15) people-dancing 16) scene-dancing
17) other-digging 18) scene-digging 19) other-drawing 20) other-drinking
21) scene-drinking 22) other-driving 23) scene-driving 24) vehicles-driving
25) other-falling 26) people-eating 27) scene-eating 28) other-falling
29) scene-falling 30) other-fighting 31) scene-fighting 32) other-flying
33) bodyparts-giving 34) scene-giving 35) other-hating 36) people-hit
37) bodyparts-holding 38) scene-holding 39) clothing-holding 40) instruments-holding
41) people-hugging 42) other-juggling 43) animals-jumping 44) vehicles-jumping
45) bodyparts-jumping 46) other-jumping 47) animals-jumping 48) vehicles-jumping
49) people-jumping 50) vehicles-jumping 51) scene-jumping 52) other-jumping
53) bodyparts-kneeling 54) people-kneeling 55) people-kissing 56) scene-kissing
57) other-kneeling 58) scene-kneeling 59) other-laughing 60) people-laughing
61) other-moaning 62) scene-moaning 63) instruments-performing 64) other-performing
65) people-performing 66) scene-performing 67) animals-playing 68) clothing-playing
69) instruments-playing 70) other-playing 71) people-playing 72) scene-playing
73) bodyparts-playing 74) bodyparts-pointing 75) other-pointing 76) people-pointing
77) scene-pointing 78) bodyparts-pointing 79) clothing-pointing 80) other-pointing
81) people-poising 82) scene-poising 83) other-poising 84) people-poising
85) vehicles-pushing 86) other-reaching 87) scene-reaching 88) other-reaching
89) people-reaching 90) animals-reaching 91) other-riding 92) people-riding
93) bodyparts-sitting 94) vehicles-sitting 95) animals-running 96) bodyparts-running
97) clothing-running 98) scene-running 99) people-running 100) scene-running
101) vehicles-running 102) other-serving 103) people-serving 104) other-serving
105) instruments-serving 106) other-serving 107) people-serving 108) animals-serving
109) bodyparts-serving 110) clothing-serving 111) instruments-serving 112) other-serving
113) people-serving 114) scene-serving 115) vehicles-serving 116) scene-serving
117) bodyparts-swimming 118) bodyparts-swimming 119) bodyparts-swimming 120) scene-swimming
121) other-swimming 122) scene-swimming 123) bodyparts-swimming 124) clothing-swimming
125) bodyparts-swimming 126) scene-swimming 127) bodyparts-swimming 128) clothing-swimming
129) scene-swimming 130) bodyparts-swimming 131) bodyparts-swimming 132) clothing-swimming
133) other-swimming 134) scene-swimming 135) bodyparts-swimming 136) clothing-swimming
137) scene-surfing 138) bodyparts-surfing 139) bodyparts-surfing 140) clothing-surfing
141) bodyparts-surfing 142) scene-surfing 143) bodyparts-surfing 144) clothing-surfing
145) people-talking 146) scene-talking 147) other-talking 148) people-talking
149) bodyparts-talking 150) bodyparts-talking 151) other-talking 152) animals-talking
153) bodyparts-talking 154) scene-talking 155) bodyparts-talking 156) people-talking
157) bodyparts-talking 158) bodyparts-talking 159) bodyparts-talking 160) other-talking
161) bodyparts-talking 162) scene-talking 163) other-talking 164) people-talking
165) bodyparts-talking 166) bodyparts-talking 167) other-talking 168) people-talking
169) bodyparts-talking 170) bodyparts-talking 171) other-talking 172) people-talking
173) bodyparts-talking 174) bodyparts-talking 175) other-talking 176) people-talking
177) bodyparts-talking 178) bodyparts-talking 179) other-talking 180) people-talking
181) climbing 182) climbing 183) climbing 184) climbing
185) swimming 186) swimming 187) swimming 188) swimming
189) swimming 190) swimming 191) swimming 192) swimming
193) cooking 194) cooking 195) cooking 196) cooking
197) swinging 198) swinging 199) swinging 200) swinging
201) fighting 202) fighting 203) fighting 204) fighting
205) pointing 206) pointing 207) pointing 208) pointing
209) fishing 210) fishing 211) fishing 212) fishing
213) kneeling 214) kneeling 215) kneeling 216) kneeling
217) shopping 218) shopping 219) shopping 220) shopping
221) falling 222) falling 223) falling 224) falling
225) touches
### C. List of phrase-pair relationships from Flickr30k Entities

#### C.1 Verbs

| Phrase-Pair | Description |
|-------------|-------------|
| 1. dog-catching-frisbee | 2. dog-holding-stick | 3. dog-jumping-ball | 4. dog-jumping-frisbee |
| 5. dog-jumping-hurdle | 6. dog-jumping-people | 7. dog-jumping-water | 8. dog-playing-ball |
| 9. dog-running-beach | 10. dog-running-field | 11. dog-running-grass | 12. dog-running-snow |
| 13. dog-running-water | 14. dog-swimming-water | 15. dogs-playing-grass | 16. dogs-playing-snow |
| 17. dogs-running-field | 18. dogs-running-grass | 19. people-blowing-bubbles | 20. people-catching-ball |
| 21. people-catching-wave | 22. people-cleaning-dishes | 23. people-climbing-mountain | 24. people-climbing-rock |
| 25. people-climbing-rock+wall | 26. people-climbing-rocks | 27. people-climbing-tree | 28. people-climbing-wall |
| 29. people-cooking-food | 30. people-cutting-cake | 31. people-dancing-people | 32. people-dancing-stage |
| 33. people-digging-snow | 34. people-drinking-beer | 35. people-eating-food | 36. people-eating-meal |
| 37. people-eating-table | 38. people-hitting-ball | 39. people-hit-tennis+ball | 40. people-holding-ball |
| 41. people-holding-book | 42. people-holding-box | 43. people-holding-camera | 44. people-holding-cup |
| 45. people-holding-dog | 46. people-holding-drink | 47. people-holding-flag | 48. people-holding-flags |
| 49. people-holding-flowers | 50. people-holding-football | 51. people-holding-guitar | 52. people-holding-microphone |
| 53. people-holding-object | 54. people-holding-people | 55. people-holding-rope | 56. people-holding-shovel |
| 57. people-holding-sign | 58. people-holding-signs | 59. people-holding-something | 60. people-holding-stick |
| 61. people-holding-tennis+shirt | 62. people-hugging-people | 63. people-jumping-ball | 64. people-jumping-bed |
| 65. people-jumping-bike | 66. people-jumping-hurdle | 67. people-jumping-people | 68. people-jumping-pool |
| 69. people-jumping-ropes | 70. people-jumping-rock | 71. people-jumping-swimming+pool | 72. people-jumping-trampoline |
| 73. people-jumping-water | 74. people-kicking-ball | 75. people-kicking-people | 76. people-kicking-soccer+ball |
| 77. people-kissing-people | 78. people-laughing-people | 79. people-painting-picture | 80. people-performing-people |
| 81. people-performing-stage | 82. people-playing-accordion | 83. people-playing-bagpipes | 84. people-playing-ball |
| 85. people-playing-baseball | 86. people-playing-basketball | 87. people-playing-board+game | 88. people-playing-psychological-game |
| 89. people-playing-dog | 90. people-playing-drum | 91. people-playing-drum + game | 92. people-playing-flute |
| 93. people-playing-football | 94. people-playing-fountain | 95. people-playing-frisbee | 96. people-playing-game |
| 97. people-playing-guitar | 98. people-playing-guitars | 99. people-playing-instrument | 100. people-playing-instruments |
| 101. people-playing-keyboard | 102. people-playing-people | 103. people-playing-piano | 104. people-playing-pool |
| 105. people-playing-sand | 106. people-playing-saxophone | 107. people-playing-snow | 108. people-playing-soccer |
| 109. people-playing-stage | 110. people-playing-swing | 111. people-playing-toy | 112. people-playing-toys |
| 113. people-playing-trumpet | 114. people-playing-violin | 115. people-playing-volleyball | 116. people-playing-water |
| 117. people-posting-news | 118. people-posting-picture | 119. people-posting-story | 120. people-posting-person |
| 121. people-posting-video | 122. people-reading-book | 123. people-reading-magazine | 124. people-reading-newspaper |
| 125. people-reading-paper | 126. people-riding-bicycle | 127. people-riding-bicycles | 128. people-riding-bike |
| 129. people-riding-bikes | 130. people-riding-bull | 131. people-riding-dirt+bike | 132. people-riding-horse |
| 133. people-riding-horses | 134. people-riding-motorbike | 135. people-riding-motorcycle | 136. people-riding-people |
| 137. people-riding-scooter | 138. people-riding-skateboard | 139. people-riding-street | 140. people-riding-surfboard |
| 141. people-riding-unicycle | 142. people-riding-wave | 143. people-running-ball | 144. people-running-beach |
| 145. people-running-field | 146. people-running-grass | 147. people-running-people | 148. people-running-road |
| 149. people-running-sidewalk | 150. people-running-street | 151. people-running-track | 152. people-running-water |
| 153. people-serving-food | 154. people-singing-guitar | 155. people-singing-microphone | 156. people-singing-people |
| 157. people-sitting-beach | 158. people-sitting-bed | 159. people-sitting-bench | 160. people-sitting-benches |
| 161. people-sitting-bike | 162. people-sitting-blanket | 163. people-sitting-boat | 164. people-sitting-building |
| 165. people-sitting-chair | 166. people-sitting-chairs | 167. people-sitting-couch | 168. people-sitting-curb |
| 169. people-sitting-desk | 170. people-sitting-dog | 171. people-sitting-floor | 172. people-sitting-grass |
| 173. people-sitting-horse | 174. people-sitting-ledge | 175. people-sitting-motorcycle | 176. people-sitting-park+bench |
| 177. people-sitting-people | 178. people-sitting-rock | 179. people-sitting-rocks | 180. people-sitting-sidewalk |
| 181. people-sitting-steps | 182. people-sitting-stool | 183. people-sitting-street | 184. people-sitting-swing |
| 185. people-sitting-table | 186. people-sitting-tables | 187. people-sitting-tree | 188. people-sitting-wall |
| 189. people-sitting-water | 190. people-sleeping-bench | 191. people-sleeping-chair | 192. people-sleeping-couch |
| 193. people-sleeping-grass | 194. people-sleeping-people | 195. people-sliding-base | 196. people-sliding-slide |
| 197. people-smiling-people | 198. people-smoking-cigarette | 199. people-standing-beach | 200. people-standing-boat |
| 201. people-standing-bridge | 202. people-standing-building | 203. people-standing-car | 204. people-standing-counter |
| 205. people-standing-door | 206. people-standing-doorway | 207. people-standing-fence | 208. people-standing-field |
| 209. people-standing-grass | 210. people-standing-ladder | 211. people-standing-line | 212. people-standing-people |
| 213. people-standing-platform | 214. people-standing-podium | 215. people-standing-road | 216. people-standing-rock |
| 217. people-standing-rocks | 218. people-standing-sidewalk | 219. people-standing-sign | 220. people-standing-snow |
| 221. people-standing-stage | 222. people-standing-street | 223. people-standing-table | 224. people-standing-tree |
| 225. people-playing-baseball | 226. people-playing-beach | 227. people-playing-wave | 228. people-playing-pool |
| 229. people-swinging-bat | 230. people-swinging-swing | 231. people-talking-cellphone | 232. people-talking-microphone |
| 233. people-talking-people | 234. people-talking-phone | 235. people-throwing-ball | 236. people-throwing-frisbee |
| 237. people-throwing-people | 238. people-walking-beach | 239. people-walking-bicycle | 240. people-walking-bike |
| 241. people-walking-bridge | 242. people-walking-building | 243. people-walking-city+street | 244. people-walking-dog |
| 245. people-walking-dogs | 246. people-walking-field | 247. people-walking-grass | 248. people-walking-hill |
| 249. people-walking-path | 250. people-walking-people | 251. people-walking-road | 252. people-walking-sidewalk |
| 253. people-walking-snow | 254. people-walking-stairs | 255. people-walking-street | 256. people-walking-trail |
| 257. people-walking-wall | 258. people-walking-water | 259. people-working-machine | 260. people-working-people |
C.2. Prepositions

1) ball-in-mouth 2) bicycle-on-street 3) boat-in-water 4) building-in-people
5) dog-in-ball 6) dog-in-collar 7) dog-in-dog 8) dog-in-field
9) dog-in-grass 10) dog-in-snow 11) dog-in-stick 12) dog-in-toy
13) dog-in-water 14) dog-on-beach 15) dog-on-grass 16) dog-on-hind+legs
17) dog-on-leash 18) dogs-in-dogs 19) dogs-in-field 20) dogs-in-grass
21) dogs-in-snow 22) dogs-in-water 23) dogs-on-grass 24) guitar-in-people
25) hands-in-people 26) object-in-mouth 27) one-in-shirt 28) other-in-shirt
29) people-across-street 30) people-behind-building 31) people-behind-counter 32) people-behind-fence
33) people-behind-people 34) people-between-people 35) people-in-area 36) people-in-back
37) people-in-ball 38) people-in-bed 39) people-in-bicycle 40) people-in-bike
41) people-in-blanket 42) people-in-boat 43) people-in-body+water 44) people-in-building
45) people-in-camera 46) people-in-cane 47) people-in-canoe 48) people-in-car
49) people-in-cart 50) people-in-chair 51) people-in-chairs 52) people-in-cigarette
53) people-in-colors 54) people-in-dirt 55) people-in-dog 56) people-in-dogs
57) people-in-doorway 58) people-in-face+paint 59) people-in-field 60) people-in-flowers
61) people-in-football 62) people-in-fountain 63) people-in-gear 64) people-in-grass
65) people-in-guitar 66) people-in-highchair 67) people-in-instruments 68) people-in-kayak
69) people-in-kitchen 70) people-in-lake 71) people-in-line 72) people-in-microphone
73) people-in-mirror 74) people-in-mud 75) people-in-number 76) people-in-ocean
77) people-in-park 78) people-in-people 79) people-in-pool 80) people-in-river
81) people-in-room 82) people-in-sand 83) people-in-snow 84) people-in-soccer+ball
85) people-in-street 86) people-in-stroller 87) people-in-swimming+pool 88) people-in-swing
89) people-in-towel 90) people-in-toy 91) people-in-toys 92) people-in-tree
93) people-in-tub 94) people-in-water 95) people-in-wheelchair 96) people-in-yard
97) people-near-beach 98) people-near-brick+wall 99) people-near-building 100) people-near-car
101) people-near-fence 102) people-near-fountain 103) people-near-lake 104) people-near-people
105) people-near-pole 106) people-near-road 107) people-near-sidewalk 108) people-near-street
109) people-near-table 110) people-near-tree 111) people-near-wall 112) people-near-water
113) people-near-window 114) people-on-back 115) people-on-balcony 116) people-on-beach
117) people-on-bed 118) people-on-bench 119) people-on-benches 120) people-on-bicycle
121) people-on-bicycles 122) people-on-bike 123) people-on-bikes 124) people-on-blanket
125) people-on-board 126) people-on-boat 127) people-on-bridge 128) people-on-building
129) people-on-bus 130) people-on-cellphone 131) people-on-chair 132) people-on-chairs
133) people-on-city+street 134) people-on-cliff 135) people-on-computer 136) people-on-couch
137) people-on-curb 138) people-on-deck 139) people-on-dock 140) people-on-fence
141) people-on-field 142) people-on-floor 143) people-on-grass 144) people-on-grill
145) people-on-hill 146) people-on-horse 147) people-on-horses 148) people-on-ice
149) people-on-ladder 150) people-on-lawn 151) people-on-ledge 152) people-on-machine
153) people-on-mat 154) people-on-motorcycle 155) people-on-motorcycles 156) people-on-mountain
157) people-on-park+bench 158) people-on-path 159) people-on-pavement 160) people-on-people
161) people-on-phone 162) people-on-pier 163) people-on-platform 164) people-on-porch
165) people-on-raft 166) people-on-rail 167) people-on-railing 168) people-on-ramp
169) people-on-road 170) people-on-rock 171) people-on-rocks 172) people-on-roof
173) people-on-rope 174) people-on-sand 175) people-on-scaffold 176) people-on-scaffolding
177) people-on-scooter 178) people-on-shore 179) people-on-side+road 180) people-on-sidewalk
181) people-on-ski-board 182) people-on-sled 183) people-on-slide 184) people-on-snowboard
185) people-on-soccer+field 186) people-on-sofa 187) people-on-stage 188) people-on-stairs
189) people-on-step 190) people-on-steps 191) people-on-stilts 192) people-on-stool
193) people-on-street 194) people-on-surfboard 195) people-on-swing 196) people-on-table
197) people-on-tire+swing 198) people-on-track 199) people-on-trail 200) people-on-train
201) people-on-trampoline 202) people-on-tree 203) people-on-walkway 204) people-on-wall
205) people-on-water 206) people-on-wave 207) people-under-tree 208) shirt-in-people
209) something-in-mouth 210) stick-in-mouth 211) street-in-people 212) table-in-people
213) tattoo-on-people 214) tennis+ball-in-mouth 215) toy-in-mouth 216) wall-in-graffiti
C.3. Clothing and Body Part Attachment

1) people-apron 2) people-aprons 3) people-arms 4) people-attire
5) people-backpack 6) people-backpacks 7) people-bag 8) people-bags
9) people-ball+cap 10) people-bandanna 11) people-baseball+cap 12) people-baseball+uniform
13) people-bathing+suit 14) people-bathing+suits 15) people-beanie 16) people-beard
17) people-beret 18) people-bikini 19) people-bikinis 20) people-black
21) people-black+jacket 22) people-black+shirt 23) people-blond-hair 24) people-blove
25) people-blue 26) people-body 27) people-boots 28) people-brown
29) people-brown+jacket 30) people-brown+shirt 31) people-business+attire 32) people-business+suit
33) people-camouflage 34) people-cap 35) people-checkered+shirt 36) people-clothes
37) people-clothing 38) people-coat 39) people-coats 40) people-collared+shirt
41) people-costume 42) people-costumes 43) people-cowboy+hat 44) people-cowboy+hats
45) people-cowboy+shirt 46) people-denim+jacket 47) people-dreadlocks 48) people-dress
49) people-dress+shirt 50) people-dresses 51) people-eyes 52) people-face
53) people-faces 54) people-feet 55) people-finger 56) people-fingers
57) people-flip-flops 58) people-garb 59) people-glasses 60) people-gloves
61) people-goggles 62) people-gold 63) people-gray 64) people-green
65) people-hair 66) people-haircut 67) people-hand 68) people-hands
69) people-head 70) people-hat 71) people-hats 72) people-head
73) people-headband 74) people-headphones 75) people-headscarf 76) people-heads
77) people-heels 78) people-helmet 79) people-helmets 80) people-hoodie
81) people-jacket 82) people-jackets 83) people-jean+shorts 84) people-jeans
85) people-jersey 86) people-jerseys 87) people-jumpsuit 88) people-khaki+pants
89) people-jumpsuit 90) people-kilts 91) people-lab+coat 92) people-lap
93) people-leather+jacket 94) people-leg 95) people-legs 96) people-leotard
97) people-life+jacket 98) people-life+jackets 99) people-make up 100) people-mask
101) people-mohawk 102) people-mouth 103) people-mustache 104) people-necklace
105) people-nose 106) people-orange 107) people-orange+dress 108) people-orange+hat
109) people-orange+shirt 110) people-orange+jacket 111) people-orange+vest 112) people-orange+vests
113) people-outfit 114) people-outfits 115) people-overalls 116) people-pajamas
117) people-pants 118) people-pants 119) people-pigtails 120) people-pink
121) people-pink+coat 122) people-pink+dress 123) people-pink+hat 124) people-pink+jacket
125) people-pink+shirt 126) people-pink+pants 127) people-pink+skirt 128) people-pink+swater
129) people-plaid+shirt 130) people-poly+shirt 131) people-ponytail 132) people-purple
133) people-purple+shirt 134) people-purple+skirt 135) people-purple+socks 136) people-red+white
137) people-red+shirt 138) people-red+jacket 139) people-red+shirt 140) people-robes
141) people-rock+face 142) people-safety+vest 143) people-safety+vests 144) people-red+shirt
145) people-scarf 146) people-scrubs 147) people-shirt 148) people-shirts
149) people-shoes 150) people-shoes 151) people-shop+bag 152) people-shop+bags
153) people-shorts 154) people-shorts 155) people-shop+coat 156) people-skirt
157) people-skirts 158) people-sleeveless+shirt 159) people-smile 160) people-sneakers
161) people-snow+shoes 162) people-snowsuit 163) people-socks 164) people-straw+hat
165) people-striped+shirt 166) people-suit 167) people-suits 168) people-sunglasses
169) people-suspenders 170) people-sweater 171) people-sweathirt 172) people-svims+trunks
173) people-swimming+trunks 174) people-swimsuit 175) people-swimsuits 176) people-t-shirt
177) people-t-shirts 178) people-tan+jacket 179) people-tan+pants 180) people-tan+shirt
181) people-tan+shirt 182) people-tank 183) people-tattoo 184) people-tattoos
185) people-teeth 186) people-thumbs 187) people-tie 188) people-tongue
189) people-top 190) people-top 191) people-trunks 192) people-turban
193) people-tuxedo 194) people-umbrella 195) people-umbrellas 196) people-underwear
197) people-uniform 198) people-uniforms 199) people-vest 200) people-vests
201) people-wedding+dress 202) people-wetsuit 203) people-white 204) people-wig
205) people-winter+clothes 206) people-winter+clothing 207) people-white