Visual Relationship Detection with Multiple Cues

Arun Mallya Bryan A. Plummer Svetlana Lazebnik
University of Illinois at Urbana Champaign
{amallya2, bplumme2, slazebni}@illinois.edu

In this work, we propose a framework to solve the recently introduced task of Visual Relationship Detection (VRD) by Lu et al. [10]. Given an image, 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. The VRD dataset has a vocabulary of 100 object classes and 70 predicates annotated in 4000 training and 1000 test images.

It would seem advantageous to train 100 object detectors on this dataset, as was done by Lu et al. [10]. 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 of dimensionality 4096 from a Fast RCNN model trained on MSCOCO [9] due to the larger range of categories than PASCAL, and word2vec of dimensionality 300 for object and predicate class names.

Image-Class Compatibility Features. We train the following CCA models to measure the compatibility of a region proposal with the available classes and relationships:

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

Given a candidate subject box b, object box b', and relationship r, the concatenation of all the above CCA features gives us $\phi_{CCA}(b, b', r)$. Each candidate relationship gets six CCA scores (model 1 in the above list is applied both to the subject and the object).

Subject/Object Size Features. People have a bias towards describing larger, more salient objects, and object classes are often biased towards a certain size and scale in most images leading prior work to consider the size of a candidate box in their models [3, 7, 11]. We follow the procedure of [11], so that given a box b with dimensions normalized by the image size, we have

$$\phi_{size}(b) = 1 - b_{width} \times b_{height}.$$

Subject/Object Position Features. The location of a bounding box in an image has been shown to be predictive of the kinds of phrases it may refer to [2, 4, 7, 8]. 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 [1]. We randomly sample EdgeBox [12] proposals with IOU < 0.5 with the ground truth boxes for negative examples. Our scoring function is

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

where $\text{SVM}_{class(b)}$ returns the probability of finding a box of subject/object class $class(b)$ at the proposed location (we use Platt scaling to convert the SVM output to a probability). The encodes knowledge about the location priors of various object/subject classes in the image.

Relative Subject-Object Position. Relationships between a subject and object also constrain the relative position between them. For example, (*man*, *on*, *horse*) implies that the horse is below the man. Given a subject and object box

with coordinates $b = (x, y, w, h)$ and $b' = (x', y', w', h')$ respectively, we compute a four-dim. feature

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

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 to obtain a probability output. This is similar to the method of [6], but rather than learning a Gaussian Mixture Model using only positive data, we learn a more discriminative model. We train an SVM per relationship type (70 in all) to obtain our last feature

$$\phi_{rel\_pos}(b, b', r) = -\log(\text{SVM}_r(b, b')).$$

Thus for a relationship r with subject and object box b and b' respectively, we obtain a 11-dimensional feature: ϕ_{CCA} of length 6, ϕ_{size} of length 2 (one per subject and object), ϕ_{pos} of length 2 (one per subject and object), and ϕ_{rel_pos} of length 1. We train a linear rank-SVM model [5] 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 [10] as this allows us 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.

Phrase detection results

In the case of relationship detection, we have to output a set of (*subject*, *predicate*, *object*) relationships and localize both the *subject* and *object* boxes with an Intersection-Over-Union (IOU) of at least 0.5 with their corresponding ground truth boxes. For phrase detection, the task is to localize the entire relationship as one bounding box (union of subject and object boxes) while ensuring that this box has at least 0.5 overlap with the ground truth relationship box.

The results of phrase and relationship detection on the test set of the Stanford VRD dataset are reported in Table 1. Consistent with [10], we report recall, R@{100, 50}, or the fraction of time the correctly localized phrase/relationship was in the top 100 (resp. 50) ranked phrases/relationships in the image. The right side shows performance for phrases/relationships that have not been encountered in the training set. Our method clearly outperforms that of Lu *et al*. [10], which uses separate visual, language, and relationship likelihood cues. We also observe that cues based on object class and relative subject-object position provide a noticeable boost in performance. Further, due to our use of CCA with continuous multi-modal embeddings, we generalize better to unseen relationships.

Visualizations of detected relationships

Figure 1 shows some of the highly confident and correctly localized detections. We detect various 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*).

Figure 2 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 1, the relationship (*person*, *has*, *shorts*) was marked as present, whereas the middle, second row image in Figure 2 has (*person*, *has*, *hat*) marked as absent, which indicates a lapse in annotation and scope for improvement in the dataset.

Figure 3 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.

Conclusions

In this work, we have shown how we can use multiple cues to solve the problem of VRD, without requiring end-to-end training and complicated models. The positions of the subject and object boxes as well as their relative positions play an important role in improving performance, as shown in Table 1. Further, the use of continuous embeddings and simpler CCA models help generalization to the zero-shot case, clearly outperforming the prior work.

References

[1] 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`.

[2] 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.

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

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

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

[6] 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.

	Method	Phrase Det.		Rel. Det.		Zero-shot Phrase Det.		Zero-shot Rel. Det.	
		R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50
(a)	Visual Only Model [10]	2.61	2.24	1.85	1.58	1.12	0.95	0.78	0.67
	Visual + Language + Likelihood Model [10]	17.03	16.17	14.70	13.86	3.75	3.36	3.52	3.13
(b)	CCA	15.36	11.38	13.69	10.08	12.40	7.78	11.12	6.59
	CCA + Size	15.85	11.72	14.05	10.36	12.92	8.04	11.46	6.76
	CCA + Size + Position	**20.70**	**16.89**	**18.37**	**15.08**	**15.23**	**10.86**	**13.43**	**9.67**

Table 1. Phrase and Relationship detection recall at different thresholds (R@{100,50}). CCA refers to the combination of six CCA models. Position refers to the combination of individual box position and pairwise spatial classifiers.

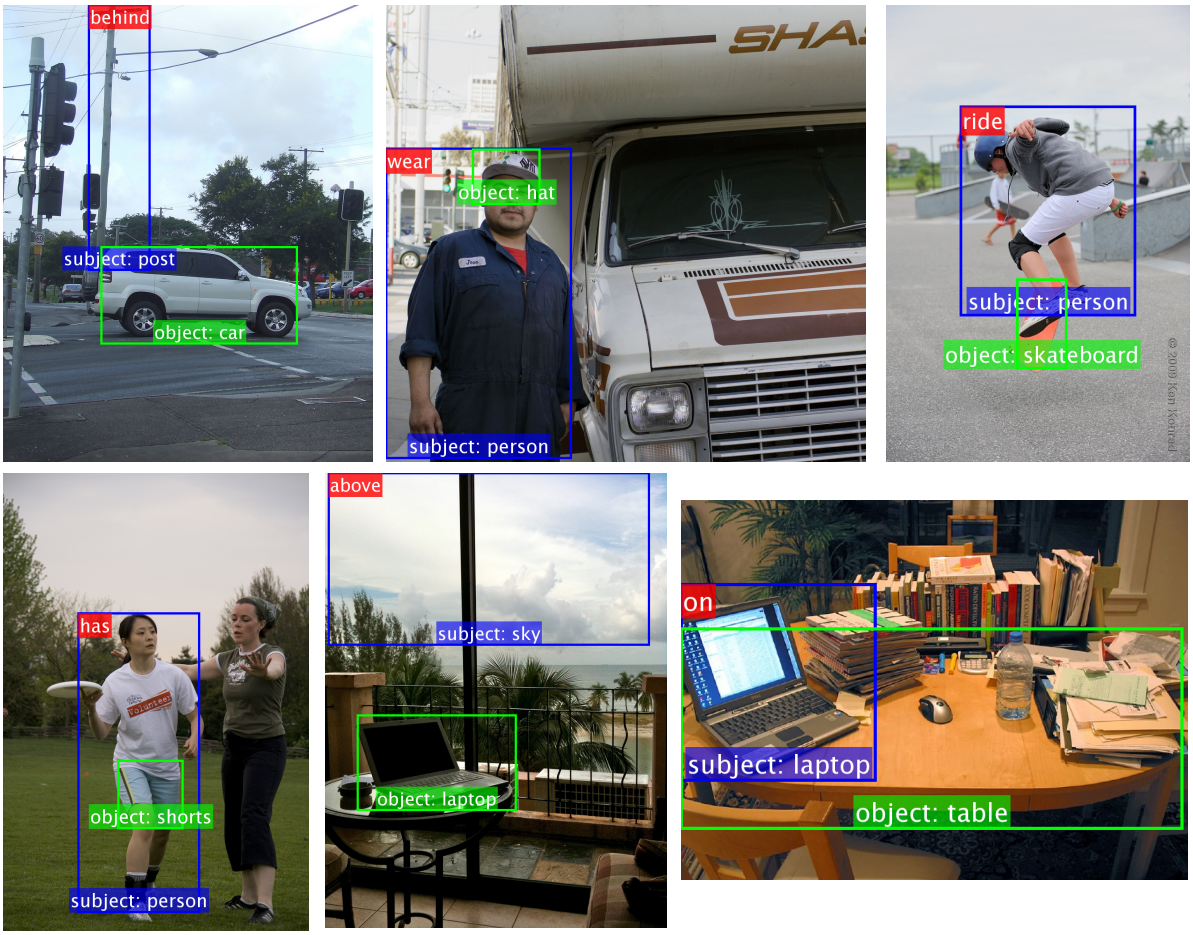

Figure 1. Highly confident and correctly localized relationships on the VRD dataset.

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

[8] 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.

[9] 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.

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

[11] 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*, 2016.

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

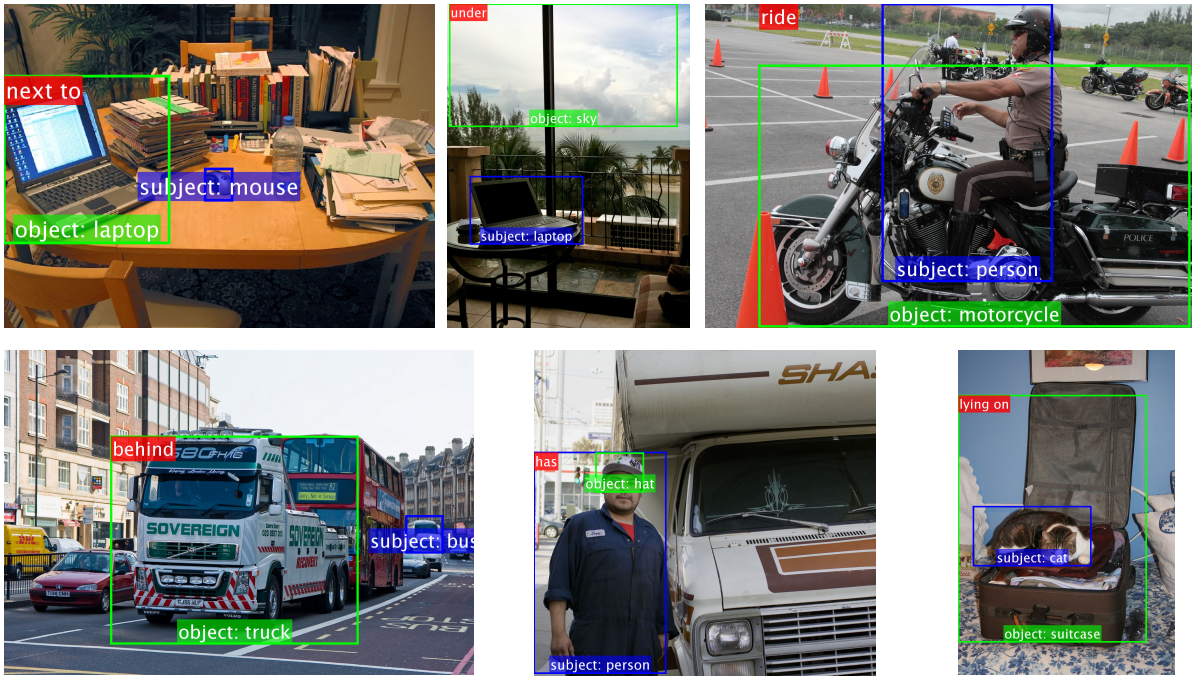

Figure 2. Plausible and logically correct detected relationships, penalized as negatives due to lack of annotations in the VRD dataset.

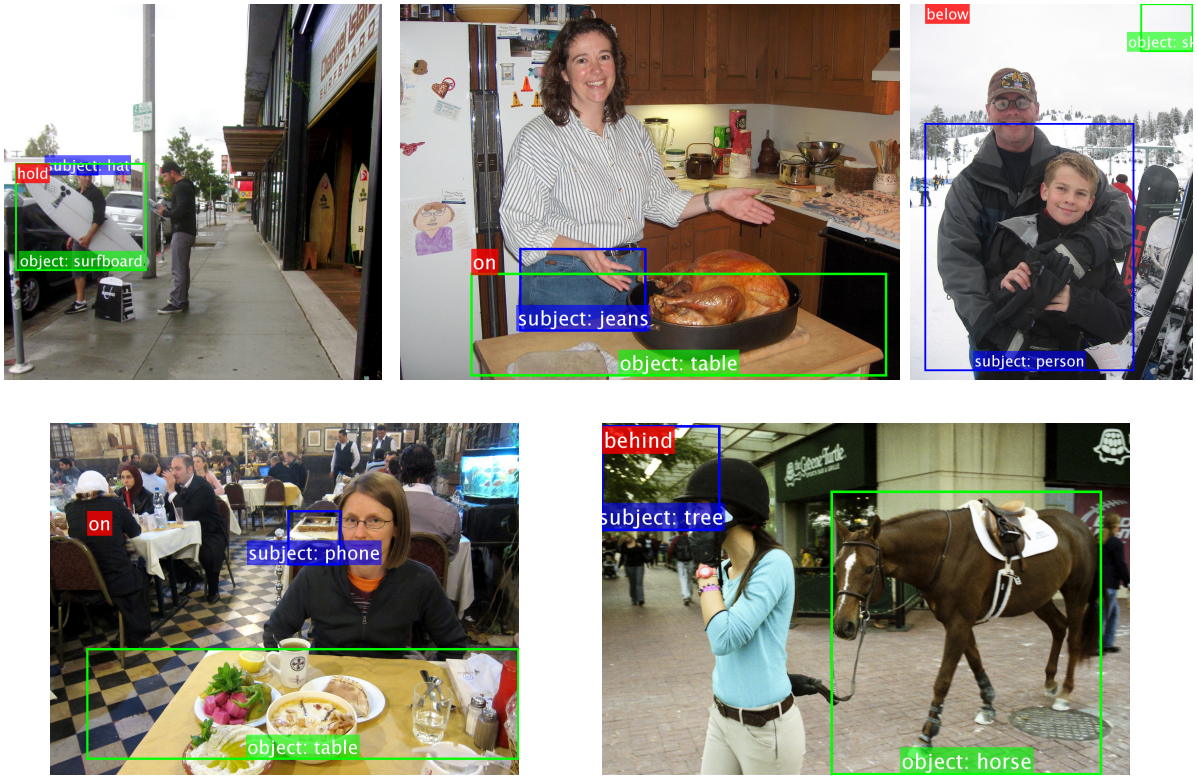

Figure 3. Falsely detected relationships on the VRD dataset. Mistakes are either due to incorrect localization of objects, prediction of implausible relationships, contextually incorrect relationships, or a combination of mistakes.