Visual Relationship Detection with Multiple Cues

Arun Mallya, Bryan A. Plummer, Svetlana Lazebnik
University of Illinois at Urbana-Champaign

vision.cs.illinois.edu/go/vrd

Overview

Visual Relationship Detection

Subject
Proposed Class: **cat**
Box: $b_s(x_s, y_s, w_s, h_s)$
Region embedding: VGG(b_s)
Class embedding: w2v(**cat**)

Object
Proposed Class: **suitcase**
Box: $b_o(x_o, y_o, w_o, h_o)$
Region embedding: VGG(b_o)
Class embedding: w2v(**suitcase**)

Predicate
Proposed Class: **lying on**
Union Box: $b_s \cup b_o$
Class embedding: w2v(**lying on**)

VGG(*b*) refers to the 4096-d fc7 features from a Fast RCNN model trained on MSCOCO
w2v(a phrase) refers to the word-averaged 300-d word2vec feature

Proposed Visual and Language Cues

1. **Image-Class Compatibility Scores (6-dim)**
   Measures compatibility of regions with proposed classes and relationships using normalized Canonical Correlation Analysis (CCA).

   Φ_{CCA}
   1. CCA (VGG(b_s), w2v(**cat**))
   2. CCA (VGG(b_o), w2v(**suitcase**))
   3. CCA (VGG(b_s), [w2v(**cat**), w2v(**lying on**)])
   4. CCA (VGG(b_o), [w2v(**lying on**), w2v(**suitcase**)])
   5. CCA (VGG($b_s \cup b_o$), w2v(**lying on**))
   6. CCA (VGG($b_s \cup b_o$), [w2v(**cat**), w2v(**lying on**), w2v(**suitcase**)])

2. **Subject/Object Size (2-dim)**
   Captures tendency of objects to be of certain size and scale in images [2].
   $\Phi_{size} = 1 - (b_{width} \times b_{height})$

3. **Subject/Object Position (2-dim)**
   Captures tendency of objects to be in certain positions in images.
   $\Phi_{pos} = -\log(\textit{position classifier SVM})$

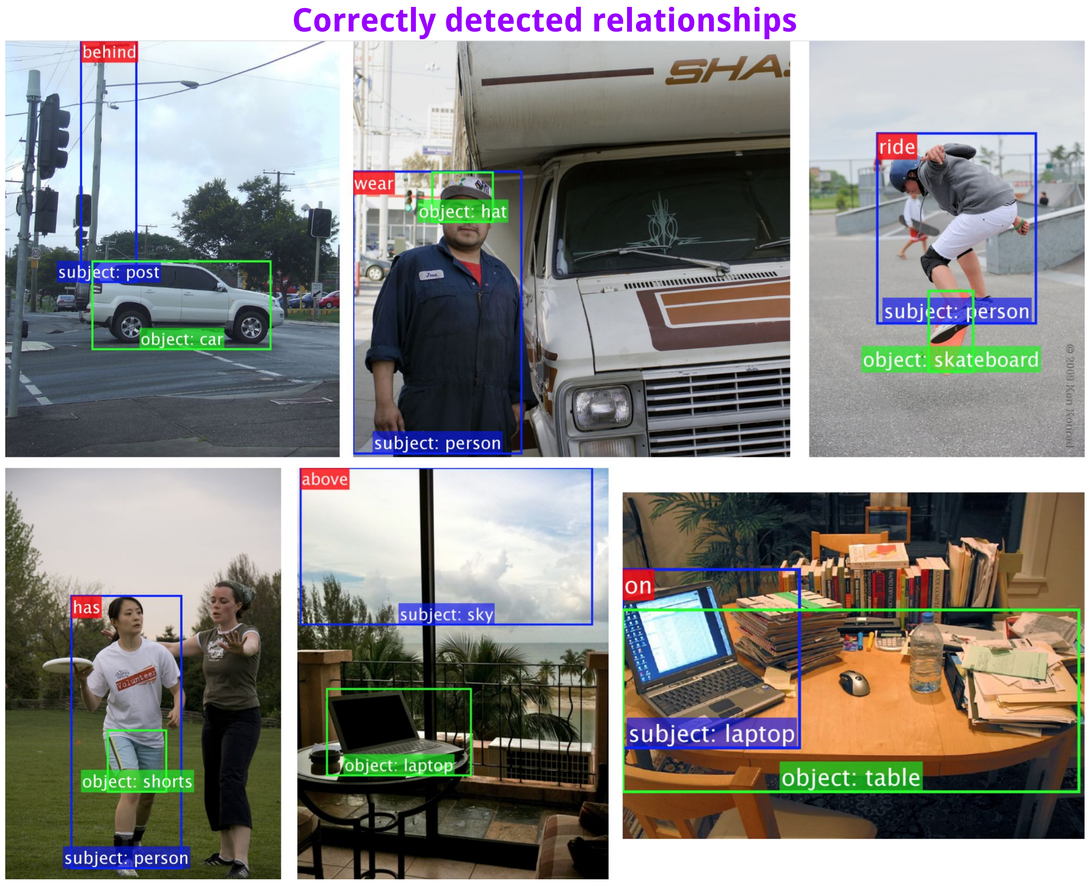
4. **Relative Subject/Object Position (1-dim)**
   Measures compatibility of relative subject-object positions with predicate [3].
   $\Phi_{rel\text{-}pos} = -\log(\textit{relative position classifier SVM})$

Final Model:

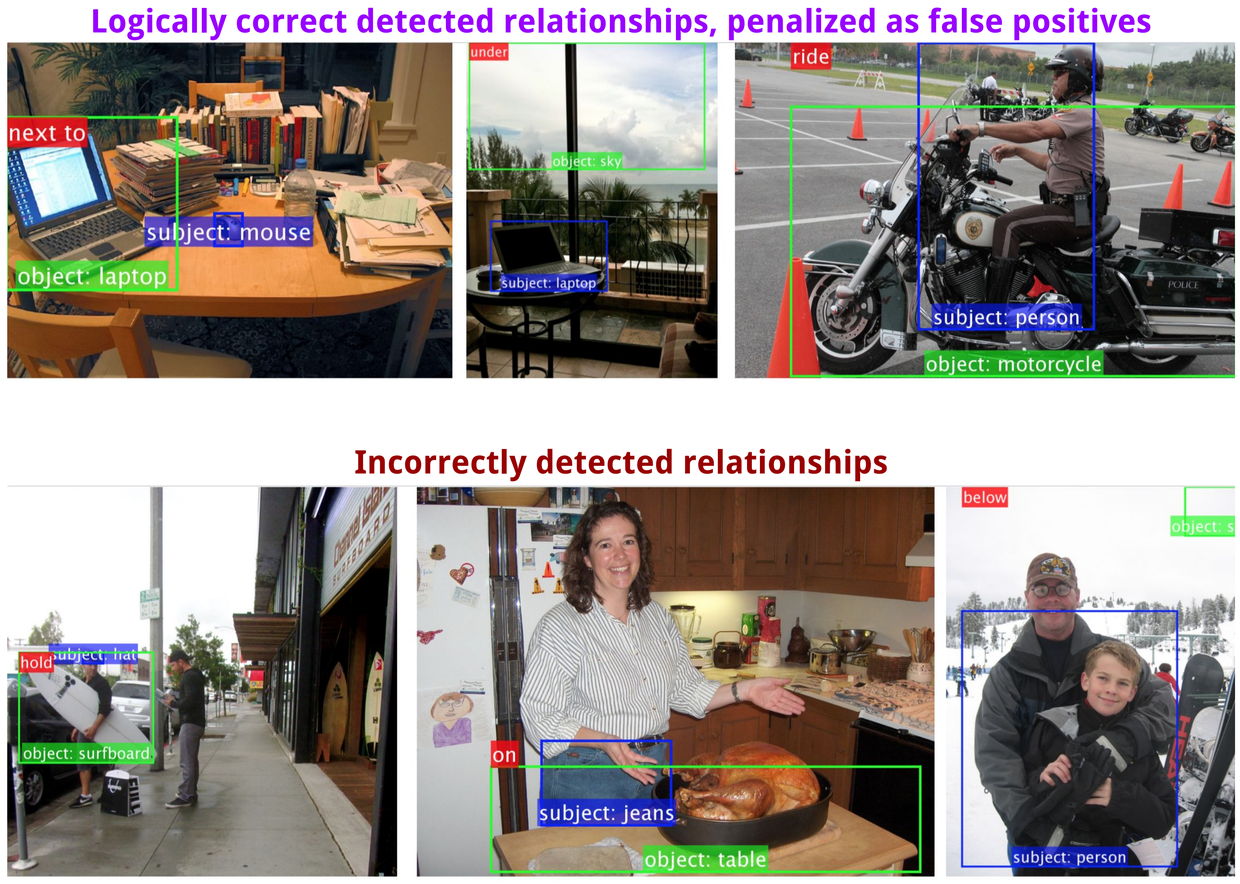
Rank-SVM on concatenated features $\Phi = [\, \Phi_{CCA}, \Phi_{size}, \Phi_{pos}, \Phi_{rel\text{-}pos} \,]$
- Negative Proposals from EdgeBoxes
- Test-time proposals provided by detectors trained in [1]

Results

Correctly detected relationships



Logically correct detected relationships, penalized as false positives



Incorrectly detected relationships

Phrase and Relationship detection recall at different thresholds on the VRD Dataset [1].

Method	Phrase Detection		Relationship Detection		Zero-shot Phrase Detection		Zero-shot Relationship Detection	
	Recall @50	Recall @100	Recall @50	Recall @100	Recall @50	Recall @100	Recall @50	Recall @100
Visual-Only Model [1]	2.24	2.61	1.58	1.85	0.95	1.12	0.67	0.78
Visual + Language + Likelihood Model [1]	16.17	17.03	13.86	14.70	3.36	3.75	3.13	3.52
CCA	11.38	15.36	10.08	13.69	7.78	12.40	6.59	11.12
CCA + Size	11.72	15.85	10.36	14.05	8.04	12.92	6.76	11.46
CCA + Size + Position	**16.89**	**20.70**	**15.08**	**18.37**	**10.86**	**15.23**	**9.67**	**13.43**

Key Takeaways:

- Competitive or better performance than previous methods, without the use of end-to-end training and complicated models.
- Position cues greatly help improve performance as relationships constrain mutual position
- Continuous embeddings and simpler CCA models help generalization resulting in improved performance for the zero-shot setting.

References

1. C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In ECCV, 2016.
2. B. A. Plummer, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to- sentence models. IJCV, 2016.
3. 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.