Combining Geometric, Textual and Visual Features for Predicting Prepositions in Image Descriptions

Arnau Ramisa[1], Josiah Wang*[2], Ying Lu[3], Emmanuel Dellandrea[3], Francesc Moreno-Noguer[1], Robert Gaizauskas[2]*

1) Institut de Robòtica i Informàtica Industrial (UPC-CSIC), Barcelona, Spain
2) Department of Computer Science, University of Sheffield, UK
3)LIRIS, Ecole Centrale de Lyon, France

* Denotes equal contribution

Problem

- We address the prediction of a preposition linking two entities (**trajector** and **landmark**), detected in an image.
- Two cases considered: with known entity labels, and when they are determined jointly with the preposition.

Approach

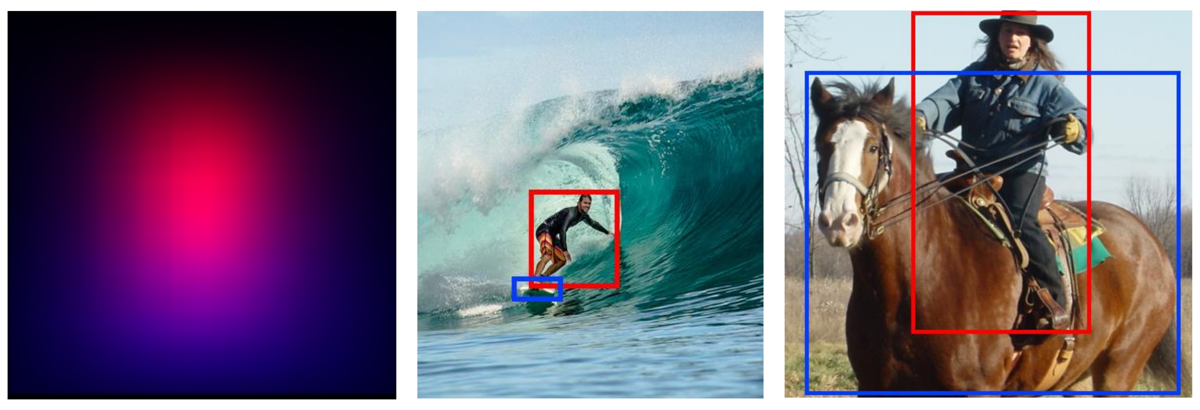
- Textual, visual and geometric features are evaluated to predict the preposition with a linear classifier (observed entity labels) and with a chain CRF (hidden entity labels).

Contributions

- The three feature types can contribute to the prediction task.
- Text embeddings add robustness against label sparsity.

Under

On



Geometric features

- Vector connecting T and L centers normalized by E (2)
  $\vec{V}$ / max(enclosing box height, width)
- Area of trajector bounding box relative to landmark (1)
  (T height × T width) / (L height × L width)
- Aspect ratio of each bounding box (2)
  (T width / T height) ; (L width / L height)
- Area of each bounding box w.r.t. enclosing box (2)
  (<T,L> height × <T,L> width) / (E height × E width)
- Intersection over union of the bounding boxes (1)
  π / { (T height × T width)+(L height × L width) - π }
- Distance between T and L, normed by picture size (1)
  $|\vec{v}|2$ / max(picture height, width)
- Area of each bounding box w.r.t. the whole image (2)
  (<T,L> height × <T,L> width) / (P height × P width)

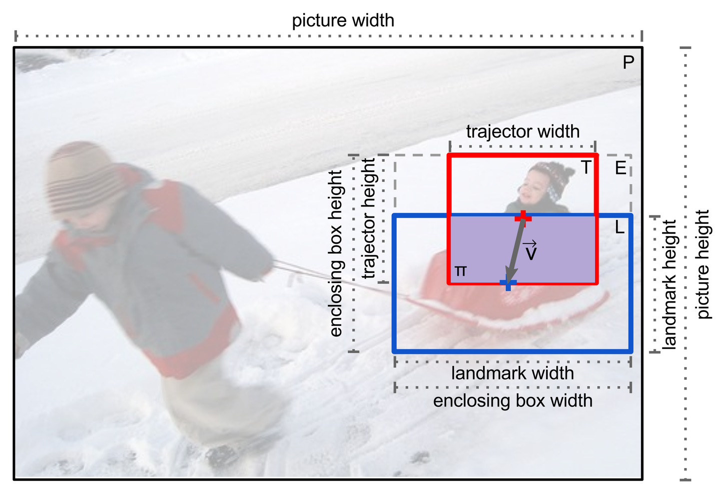
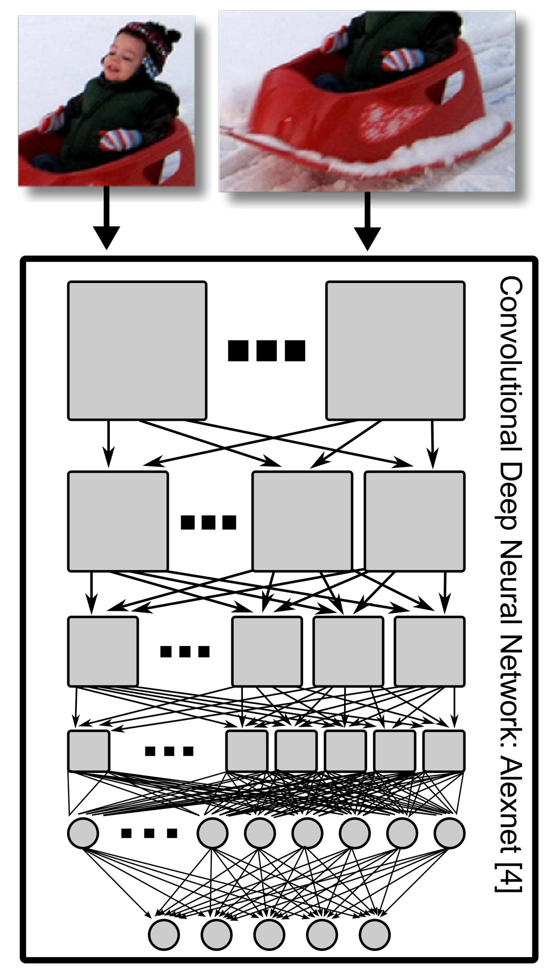
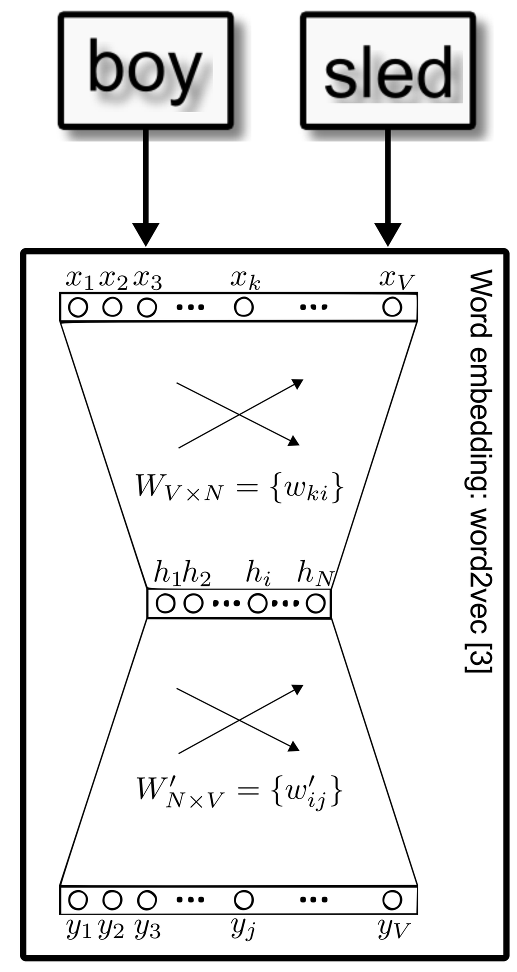
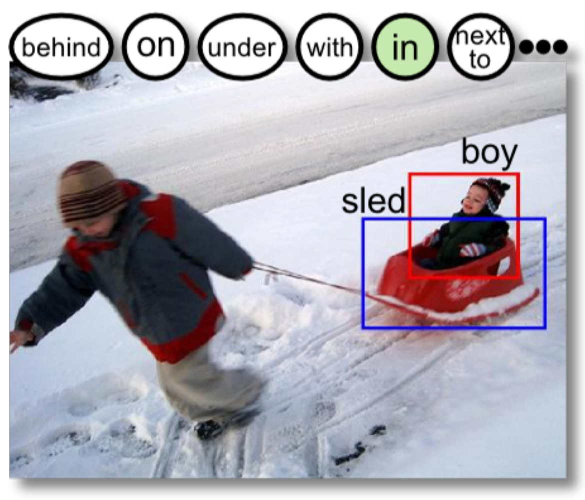


Image features

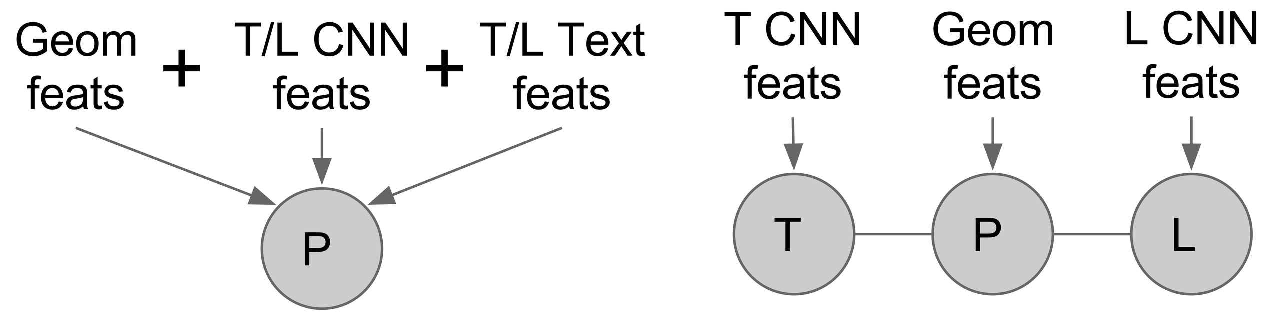


Text features





Learning models



Logistic Regression model

Chain CRF model

Datasets

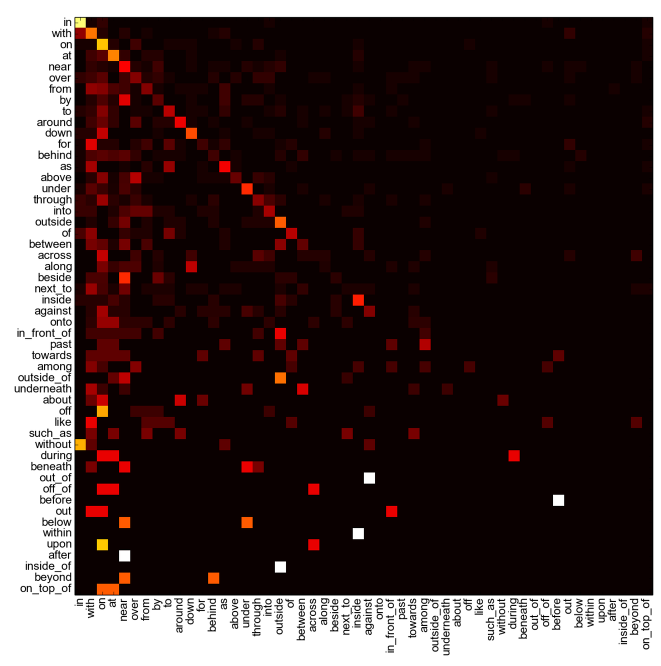
- For evaluation, we used two large-scale image datasets with human authored descriptions: MSCOCO [1] and Flickr30k [2].
- Prepositional relations relevant to the image are detected using Stanford CoreNLP, and cleaned manually.
- To avoid data sparseness in Flickr30k we extract the lemmatised head word of the original phrase using the Collins (2003) semantic head finding rules in Stanford CoreNLP.
- We consider two variants of trajector and landmark terms in our experiments:
  - Using the provided high-level categories (80 for MSCOCO and 8 for Flickr8k).
  - Using the original terms occuring in the sentence, which constitute a bigger and more realistic challenge.
- Dataset Sizes:
  MSCOCO: 8,029 training and 3,431 testing instances.
  Flickr30k: 46,847 training and 20,010 testing instances.

Evaluation

- Multiple prepositions may be suitable for a trajector-landmark pair, hence we propose to use **mean rank** as evaluation metric, but we also report accuracy for comparison purposes.
- As a baseline, we rank the prepositions by their relative frequency in the training set, which gives surprisingly good results.

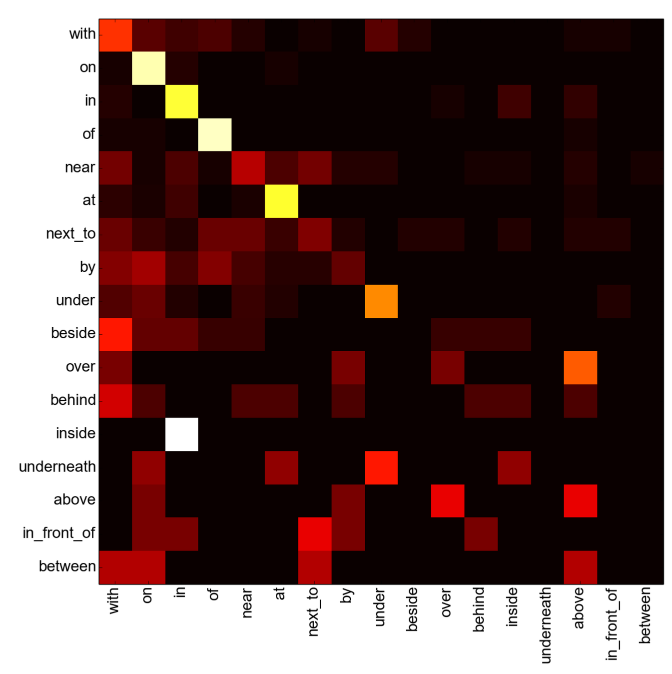
Top: Mean rank of the correct preposition (lower is better). Bottom: Accuracy with different feature configurations. All results are with the original trajector/landmark terms from descriptions. IND stands for Indicator Vectors, W2V for Word2Vec, and GF for Geometric Features.

		IND	W2V	GF	IND+GF	W2V+GF	Baseline
Mean rank	MSCOCO (max rank 17)	1.45	1.43	1.72	1.44	**1.42**	2.14
	MSCOCO (balanced)	3.20	3.10	4.60	3.00	**2.90**	5.40
	Flickr30k (max rank 52)	1.91	1.87	2.35	1.88	**1.85**	2.54
	Flickr30k (balanced)	11.10	9.04	15.55	10.23	**8.90**	15.13
Accuracy	MSCOCO	79.7%	80.3%	68.4%	79.8%	**80.4%**	40.2%
	MSCOCO (balanced)	52.5%	**54.2%**	31.5%	52.7%	53.9%	11.9%
	Flickr30k	75.4%	75.2%	58.5%	**75.8%**	75.4%	53.7%
	Flickr30k (balanced)	24.6%	25.9%	9.0%	25.2%	**26.9%**	4.0%

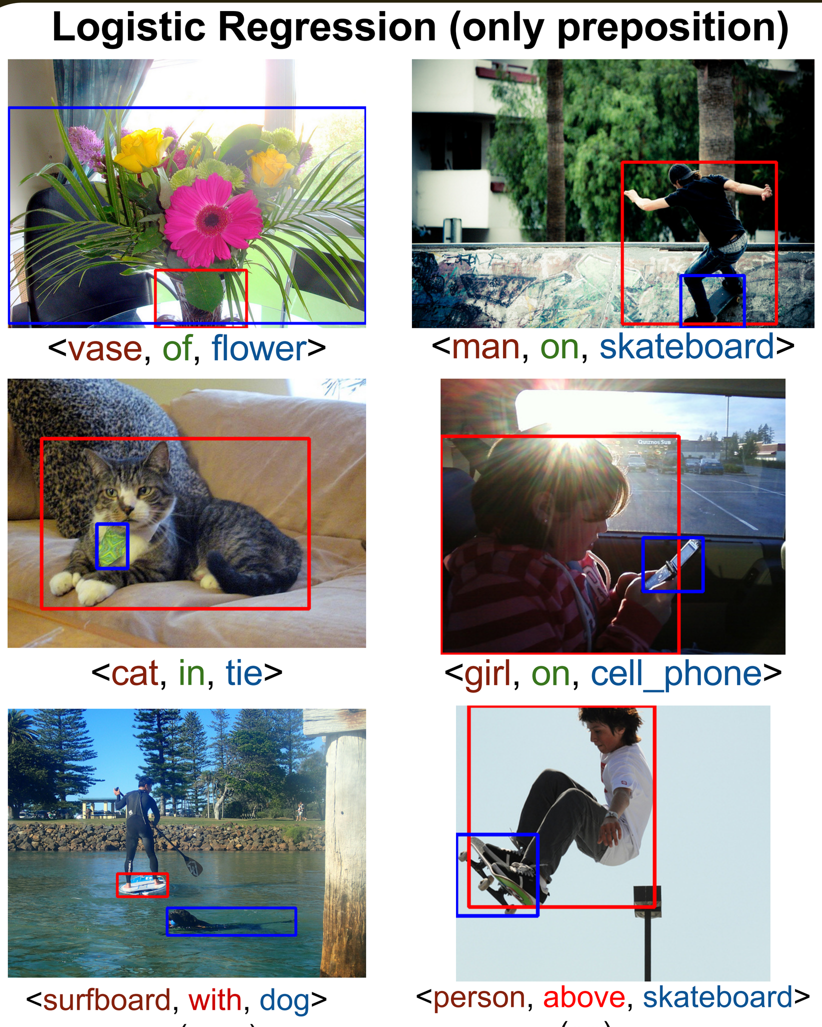


Accuracy (acc) and mean rank (rank, with max rank in parenthesis) for each variable of the CRF model, trained using the high-level concept labels. Columns under Prep (known labels) refer to the results of predicting prepositions with the trajector and landmark labels fixed to the correct values.

Dataset	Prep (known labels)		Preposition		Trajector		Landmark	
	acc	rank	acc	rank	acc	rank	acc	rank
MSCOCO	79.8%	1.46 (17)	62.9%	1.92 (17)	65.6%	4.64 (74)	44.5%	7.30 (77)
Flickr30k	67.1%	2.16 (52)	61.7%	2.28 (52)	77.3%	1.43 (8)	66.4%	1.64 (8)

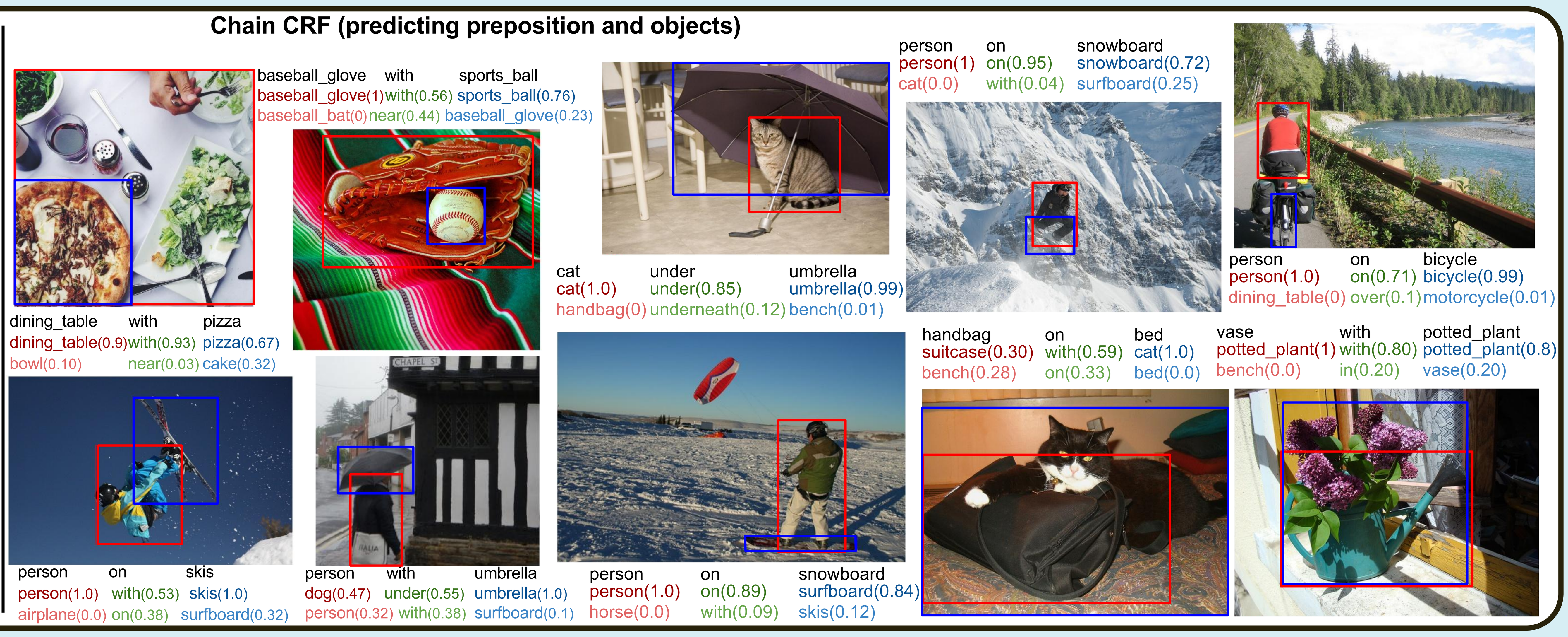


Logistic Regression (only preposition)



<vase, of, flower>
<man, on, skateboard>
<cat, in, tie>
<girl, on, cell_phone>
<surfboard, with, dog> (near)
<person, above, skateboard> (on)

Chain CRF (predicting preposition and objects)



Bibliography

[1] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Zitnick. 2014. Microsoft COCO: common objects in context. CoRR, abs/1405.0312
[2] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, S. Lazebnik. 2015. Flickr30k entities:Collecting region-to-phrase correspondences for richer image-to-sentence models.CoRR,abs/1505.04870
[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems.
[4] A. Krizhevsky, I. Sutskever, and G. Hinton. 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25, pages 1097–1105.