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

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, École Centrale de Lyon, France

{aramisa, fmoreno}@iri.upc.edu {j.k.wang, r.gaizauskas}@sheffield.ac.uk
{ying.lu, emmanuel.dellandrea}@ec-lyon.fr

Introduction

This document complements the main paper "**Combining Geometric, Textual and Visual Features for Predicting Prepositions in Image Descriptions**" (Ramisa et al., 2015). In Annex I, we detail the dataset pre-processing steps for generating instances of triples *(trajector, preposition, landmark)* for our experiments described in the main paper. In Annex II, we provide additional results from our preposition prediction experiments not included in the main paper.

Annex I: Dataset pre-processing

We base the preposition prediction task on two large-scale image datasets with human authored descriptions, namely MSCOCO and Flickr30k. Here, we discuss the pre-processing steps required to extract instances of triples $(trajector, preposition, landmark)$ from image descriptions, and to align the constituents with their corresponding bounding box instances in the image.

Flickr30k

The Flickr30k Dataset (Young et al., 2014) comprises 31,783 images with five human-authored descriptions per image. The supplementary Flickr30k Entities Dataset (Plummer et al., 2015) provides additional annotations with 244,034 cross-caption, coreference chains of mentions in descriptions of the same image that refer to the same entity. Localisation information in the form of bounding box annotations is also provided for each coreference chain entity visually depicted in its corresponding image, resulting in a total of 275,775 bounding boxes. For this paper, each entity mention that refers to multiple bounding box instances (*three women*) is treated as a single collective entity, and its bounding box expanded to be the minimal bounding box encompassing the bounding boxes of all instances of the collective entity.

To extract instances of (*trajector*, *preposition*, *landmark*) from the image descriptions, we used the transition-based dependency parser of Chen and Manning (2014) which utilises a Neural Network classifier, as implemented in Stanford CoreNLP (Manning et al., 2014). We used the Universal Dependencies representation (de Marneffe et al., 2014) which generalises prepositional terms as special case-marking elements attached to the dependent of nominal modifier relations. We extracted all such cases of *nmod*, excluding special cases like possessive and temporal modifiers. Case-markers were constrained to the 56 most common English prepositions[1], as well as several phrasal prepositions (*in front of*, *next to*). Dependency relations were retained where both the governor and its dependent overlap with the noun phrase mentions of distinct coreference chains, and where both mentions have corresponding bounding boxes. This results in a total 66,857 instances which we divided at random such that 70% is used for training and the remaining for testing, and such that the images in the training and test sets are disjoint.

The descriptive noun phrases describing entities may cause data sparseness issues (*the big, furry, black and white dog*). To alleviate this, we represented each entity with the lemmatised head word of each phrase, using a 'semantic head' variant of the head finding rules of Collins (2003) as implemented in Stanford CoreNLP. To further reduce data sparseness and to abate errors arising from head noun extraction, we grouped entities from the same coreference chain and denote each chain with a common head noun chosen by majority vote among the group, breaking ties by the most fre-

*A. Ramisa and J. Wang contributed equally to this work.

[1] http://www.clres.com/PrepositionAnalysisSummaryTable.htm

quent head noun in the corpus, and breaking any further ties at random.

MSCOCO

Microsoft Common Objects in Context (MSCOCO) (Lin et al., 2014) contains 82,783 training and 40,504 validation images (in their 2014 release), each with five human-authored descriptions. Bounding box annotations are provided for instances of 80 predefined categories (with a total of 886,284 instances).

The process of extracting preposition triples remains the same as in Flickr30k. However, an additional image-text correspondence step was required since MSCOCO does not provide direct correspondence between instances of visual entities in images and their mentions in image descriptions. As categories can be described in different ways (e.g. a *boat* could be referred to as a *boat, canoe, kayak, rowboat, ferry, motorboat*), we attempted to increase the recall of matching by expanding the category terms to a set of equivalence classes using WordNet. More specifically, each 80 category label could potentially be matched with (i) the label itself; (ii) its head noun (assuming last word of phrasal terms are head nouns, e.g. *sign* for *stop sign*); (iii) synonyms; (iv) hyponyms. The relevant word senses for each category were manually determined. A special case is with the *potted plant* category where the equivalent WordNet synset *pot plant* is a leaf node. Thus, we also included the *plant* synset to cover a larger range of potential plants such as flowers, roses and tulips. We also included the lemma *people* to the *person* category to cover cases to which people are being referred in its plural form.

Regular expression matching was performed on the image descriptions for all categories that contain at least one instance in the image. We retain triples where both trajector and landmark are nouns and where they do not refer to the same category. In addition, we also discard cases of multiple instances to which we cannot automatically determine the correct correspondence (*a person* when there are two instances of person in an image). We also removed obvious image-text matching errors (e.g. *potted plant* being matched to grass or pumpkins, or *person* to a controller or monitor). The result is a total of 8,029 preposition triples for the training set and 3,915 for the validation set.

The validation set is further annotated to remove image matching or dependency parsing errors, to act as the clean test set for our experiments. We found about 88.6% to be correct, and that 11.6% are dependency parser errors (most notably PP attachment errors) or errors in the human-authored captions, and the remaining 0.8% being further image matching errors (e.g. a *bike* being matched to a motorcycle when it actually refers to a bicycle in the image). Our final test set consists of 3,431 preposition triples.

Annex II: Additional results

The additional results included in this supplementary material are as follows:

- Figures 1 and 2 show the normalised confusion matrices for the preposition prediction task on the standard and balanced test sets respectively, using word2vec and geometric features, and the original terms from the descriptions with the logistic regression model. Notice that in the standard test set (with biased classifiers), the model often predicts the most frequent prepositions (*with*, *on* and *in*). In contrast, the predictions are much more evenly spread in the balanced test set, often with ‘reasonable’ confusions.

- Tables 1 and 2 show the per-preposition mean rank for the standard and balanced test sets respectively, using word2vec, geometric features and the original terms from the descriptions with the logistic regression model.

- Table 3 shows the mean rank and accuracy results obtained with the logistic regression model corresponding to Table 2 of the main paper, but for high-level concepts.

- Table 4 shows the mean rank and accuracy for the scenario with only trajector *or* landmark information, using the original terms found in the human-authored descriptions with the logistic regression model.

- Figures 3 and 4 visualise the distributions of the visual entity bounding boxes. Semi-transparent colored squares representing the trajector (red) and landmark (blue) are overlaid on a canvas according to their original position in the image.

- Figures 5 and 6 show qualitative results.

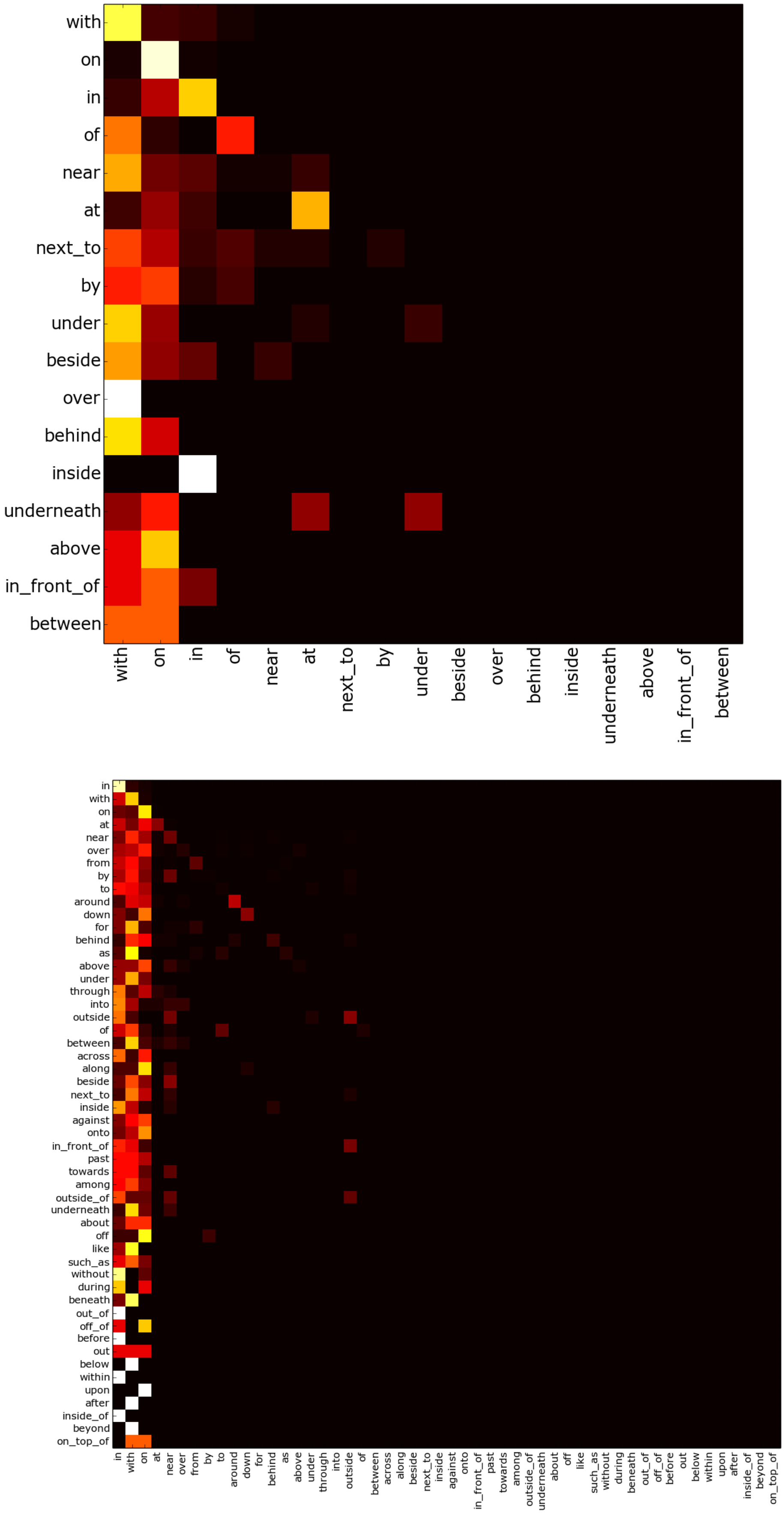


Figure 1: Normalised confusion matrices on the standard test subsets for the two datasets (top: MSCOCO, bottom: Flickr30k), using geometric features and word2vec with the original terms.

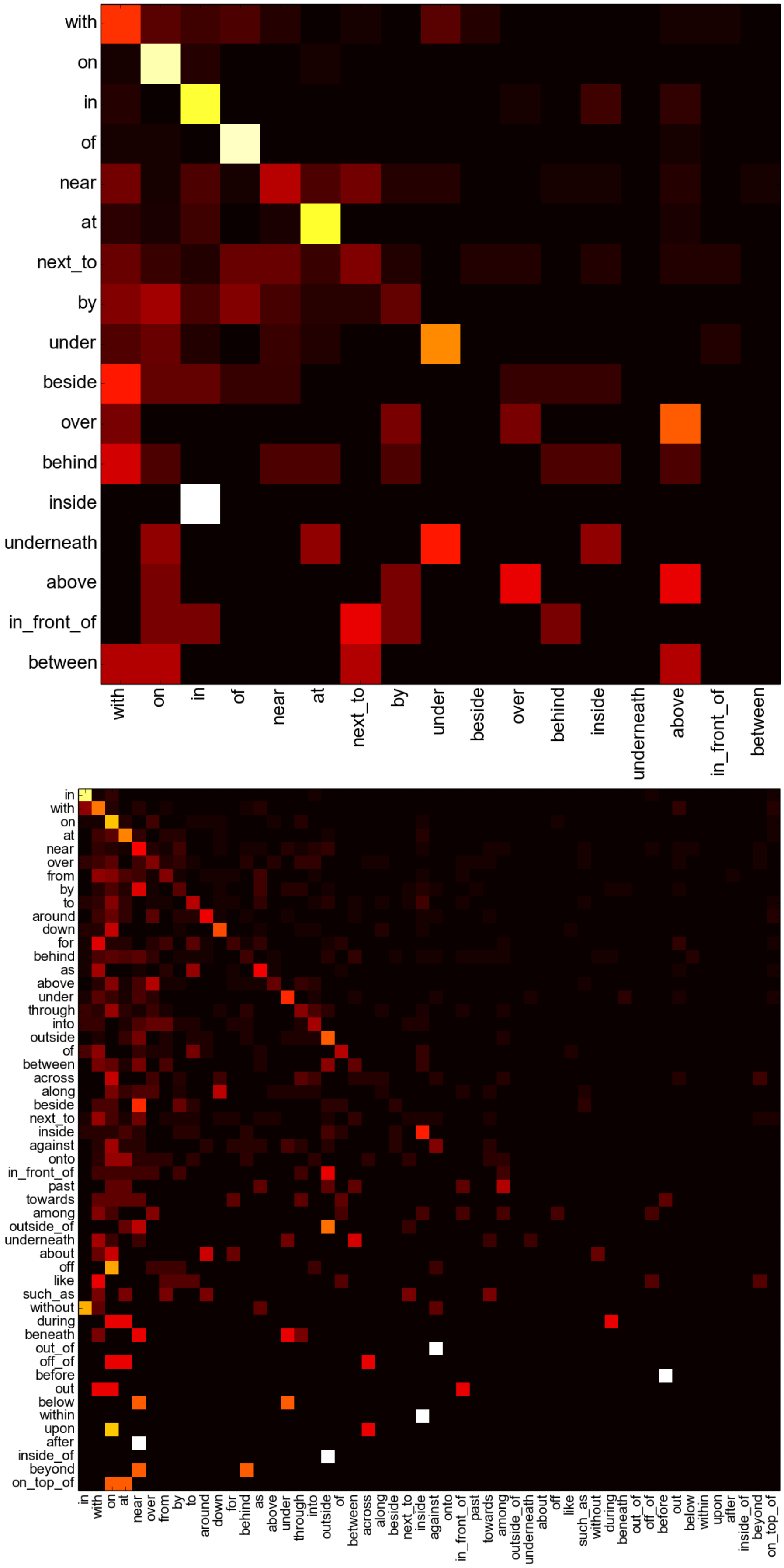


Figure 2: Normalised confusion matrices on the balanced test subsets for the two datasets (top: MSCOCO, bottom: Flickr30k), using geometric features and word2vec with the original terms.

MSCOCO (max 17)		**Flickr30k** (max 52)					
with	1.20	in	1.12	outside	8.34	like	32.78
on	1.06	with	1.39	of	7.55	such as	17.67
in	1.40	on	1.55	between	8.45	without	16.63
of	1.71	at	3.71	across	10.52	during	6.00
near	3.26	near	3.48	along	8.53	beneath	20.17
at	1.63	over	5.73	beside	11.00	out of	43.00
next to	5.68	from	6.76	next to	17.27	off of	33.00
by	5.91	by	6.36	inside	11.77	before	10.00
under	3.57	to	8.86	against	17.05	out	47.33
beside	9.27	around	6.79	onto	21.95	below	33.50
over	9.00	down	4.91	in front of	22.58	within	38.00
behind	9.30	for	7.42	past	13.00	upon	20.67
inside	15.00	behind	8.67	towards	20.50	after	26.00
underneath	6.00	as	4.74	among	23.55	inside of	28.00
above	7.50	above	9.37	outside of	10.87	beyond	39.50
in front of	12.17	under	4.93	underneath	18.38	on top of	31.00
between	11.00	through	10.14	about	21.00	**Average**	16.02
Average	6.16	into	12.03	off	29.31		

Table 1: Mean rank for each preposition using word2vec and geometric features in the standard test set with the original trajector and landmark terms.

MSCOCO (max 17)		**Flickr30k** (max 52)					
with	2.54	in	4.34	outside	10.44	like	21.67
on	1.08	with	5.44	of	5.84	such as	9.00
in	1.38	on	4.28	between	6.58	without	13.50
of	1.20	at	5.48	across	7.68	during	3.33
near	3.60	near	5.24	along	7.27	beneath	10.00
at	1.50	over	10.72	beside	10.48	out of	37.00
next to	4.61	from	11.04	next to	15.00	off of	6.33
by	5.41	by	7.72	inside	9.09	before	1.00
under	3.04	to	10.42	against	15.82	out	44.33
beside	7.33	around	9.20	onto	14.16	below	27.00
over	6.33	down	6.40	in front of	14.25	within	21.00
behind	5.70	for	7.02	past	7.25	upon	14.00
inside	15.00	behind	10.72	towards	14.88	after	19.00
underneath	3.60	as	4.28	among	16.27	inside of	8.00
above	2.50	above	11.12	outside of	6.20	beyond	13.50
in front of	7.33	under	5.14	underneath	11.31	on top of	8.50
between	8.00	through	10.57	about	9.71	**Average**	11.55
Average	4.71	into	13.57	off	18.77		

Table 2: Mean rank for each preposition using word2vec and geometric features in the balanced test set with the original trajector and landmark terms.

		IND	W2V	GF	IND+GF	W2V+GF	Baseline
Mean rank	MSCOCO (max rank 17)	1.43	1.43	1.72	**1.42**	**1.42**	2.14
	MSCOCO (balanced)	3.21	3.13	4.56	2.99	**2.94**	5.40
	Flickr30k (max rank 52)	2.24	2.24	2.35	2.13	**2.12**	2.54
	Flickr30k (balanced)	13.89	13.72	15.55	11.86	**11.77**	15.16
Accuracy	MSCOCO	80.9%	81.1%	68.4%	**81.2**%	81.1%	40.2%
	MSCOCO (balanced)	53.9%	**54.7**%	31.5%	53.7%	54.4%	11.9%
	Flickr30k	64.3%	64.3%	58.4%	**66.9**%	**66.9**%	53.7%
	Flickr30k (balanced)	12.3%	11.9%	9.0%	15.3%	**15.4**%	4.0%

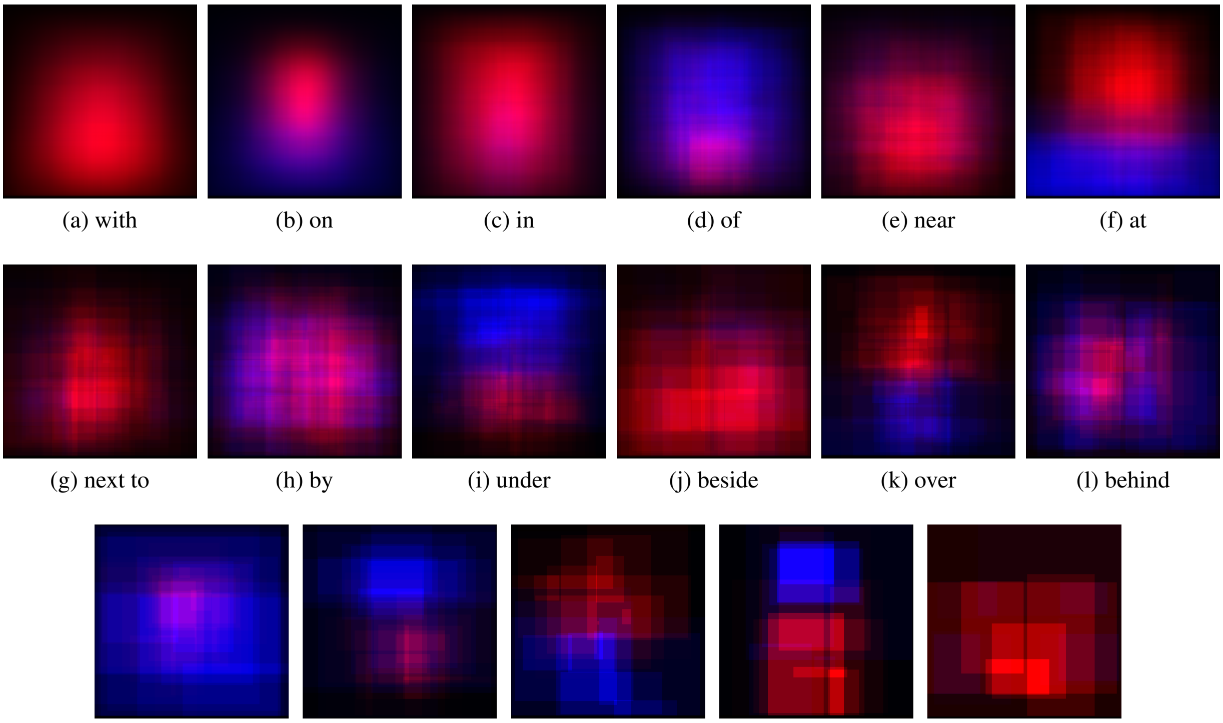
Table 3: **High-level concepts.** Top: Mean rank of the correct preposition (lower is better). Bottom: Accuracy with different feature configurations. All results are with the high-level concepts for trajector/landmark. IND stands for Indicator Vectors, W2V for Word2Vec, and GF for Geometric Features. As baseline we rank the prepositions by their relative frequencies in the training dataset.

		IND	W2V	IND+GF	W2V+GF	Baseline
Mean rank	MSCOCO (no trajector)	3.65	3.48	3.40	**3.24**	5.40
	MSCOCO (no landmark)	3.70	4.34	**3.61**	**3.61**	5.40
	Flickr30k (no trajector)	13.41	10.24	10.98	**9.77**	15.16
	Flickr30k (no landmark)	15.56	14.97	13.51	**12.72**	15.16
Accuracy	MSCOCO (no trajector)	49.4%	**51.8**%	48.7%	49.4%	11.9%
	MSCOCO (no landmark)	40.6%	35.8%	**42.0**%	41.1%	11.9%
	Flickr30k (no trajector)	25.8%	**26.1**%	22.7%	24.0%	4.0%
	Flickr30k (no landmark)	12.3%	15.1%	14.2%	**15.7**%	4.0%

Table 4: **No trajector/landmark.** Top: Mean rank of the correct preposition (lower is better). Bottom: Accuracy with different feature configurations. All results are with balanced test set, and the original concept labels for trajector/landmark, but with one of the labels withheld from the feature vector. IND stands for Indicator Vectors, W2V for Word2Vec, and GF for Geometric Features. As baseline we rank the prepositions by their relative frequencies in the training dataset.

		Mean rank (avg per class) improvement		Mean rank improvement	
		-	GF	-	GF
indicator	MSCOCO	-4.1%	-3.1%	-1.4%	-1.4%
	MSCOCO (bal)	-0.8%	-1.3%	0.0&	-1.3%
	Flickr30k	8.2%	5.5%	14.7%	11.7%
	Flickr30k (bal)	10.5%	11.5%	20.1%	13.7%
word2vec	MSCOCO	1.2%	0.2%	0.0%	0.0%
	MSCOCO (bal)	1.2%	2.2%	0.6%	1.7%
	Flickr30k	20.2%	14.7%	16.5%	12.7%
	Flickr30k (bal)	25.5%	20.5%	34.1%	24.4%

Table 5: Improvement in mean rank when going from high-level categories to the original terms used in human-authored descriptions. A larger vocabulary helps better distinguish the appropriate prepositions (especially in the case of Flickr30k which has only 8 high-level categories) at the expense of increased data sparseness, which word2vec clearly helps alleviate.

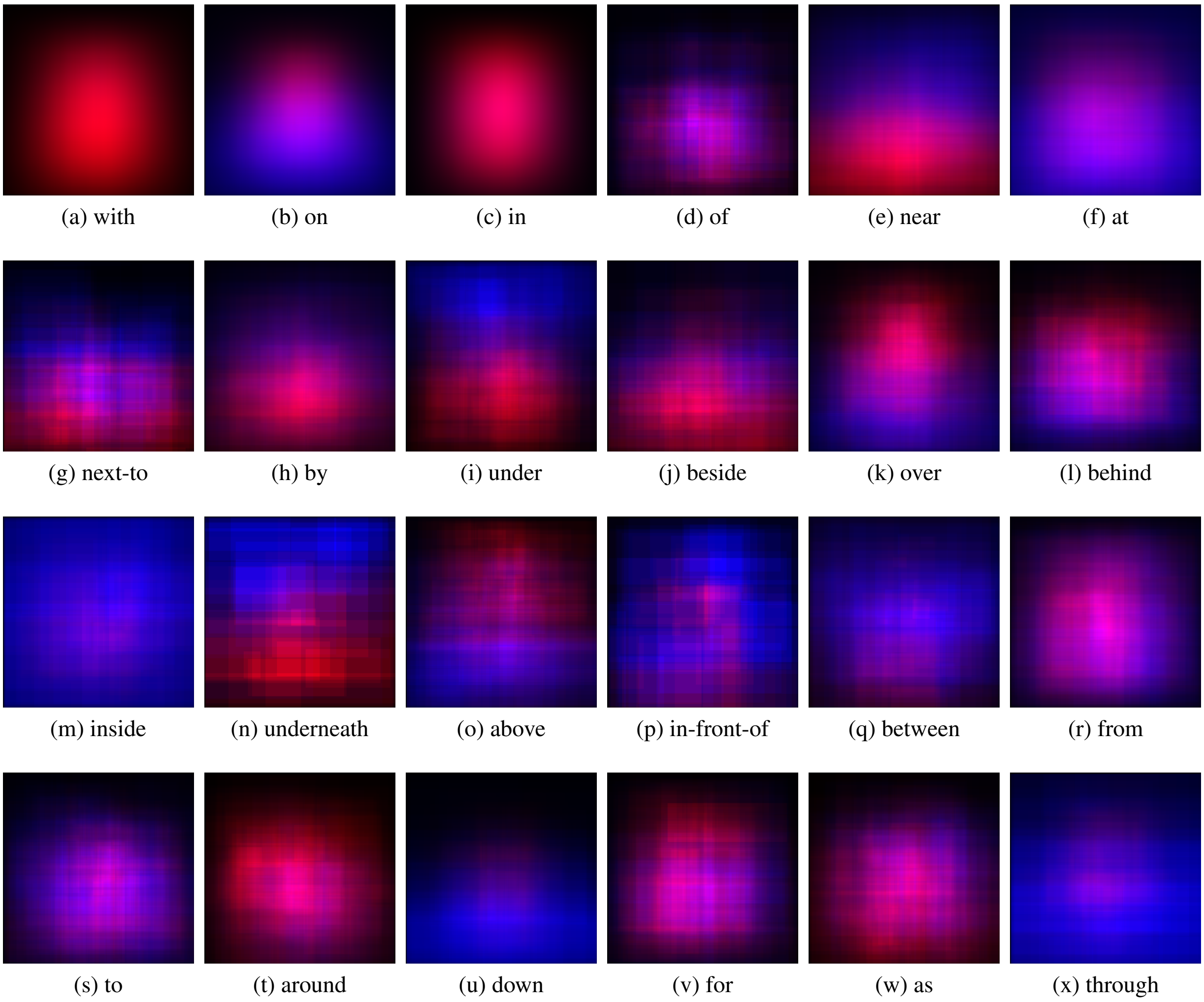


(a) with (b) on (c) in (d) of (e) near (f) at

(g) next to (h) by (i) under (j) beside (k) over (l) behind

(m) inside (n) underneath (o) above (p) in front of (q) between

Figure 3: Bounding box distributions for MSCOCO. Trajector bounding boxes are represented in red, and landmark bounding boxes in blue. Each figure shows the bounding box distribution for a specific preposition.



(a) with (b) on (c) in (d) of (e) near (f) at

(g) next-to (h) by (i) under (j) beside (k) over (l) behind

(m) inside (n) underneath (o) above (p) in-front-of (q) between (r) from

(s) to (t) around (u) down (v) for (w) as (x) through

Figure 4: Bounding box distributions for Flickr30k. Trajector bounding boxes are represented in red, and landmark bounding boxes in blue. Each figure shows the bounding box distribution for a specific preposition.

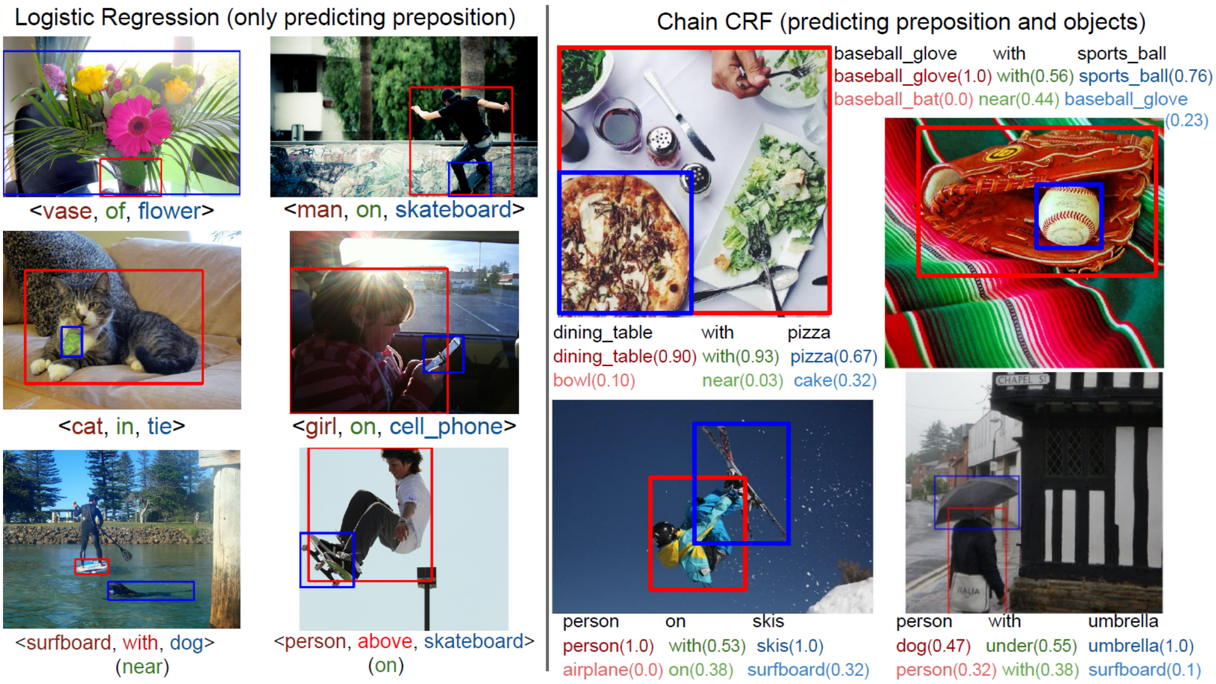

Logistic Regression (only predicting preposition)
<vase, of, flower>
<man, on, skateboard>
<cat, in, tie>
<girl, on, cell_phone>
<surfboard, with, dog>
(near)
<person, above, skateboard>
(on)
Chain CRF (predicting preposition and objects)
baseball_glove with sports_ball
baseball_glove(1.0) with(0.56) sports_ball(0.76)
baseball_bat(0.0) near(0.44) baseball_glove (0.23)
dining_table with pizza
dining_table(0.90) with(0.93) pizza(0.67)
bowl(0.10) near(0.03) cake(0.32)
person on skis
person(1.0) with(0.53) skis(1.0)
airplane(0.0) on(0.38) surfboard(0.32)
person with umbrella
dog(0.47) under(0.55) umbrella(1.0)
person(0.32) with(0.38) surfboard(0.1)

Figure 5: **Qualitative results.** Left: Example output using the logistic regression model to predict only the preposition (the correct preposition is shown in parenthesis for misclassified cases in the bottom row). Right: Example output using the chain CRF model to predict the preposition, trajector and landmark. The first row for each image shows the ground truth labels for the trajector, preposition and landmark. Subsequent rows show the predicted labels for each, stacked in order of confidence (probabilities for each label in parenthesis).

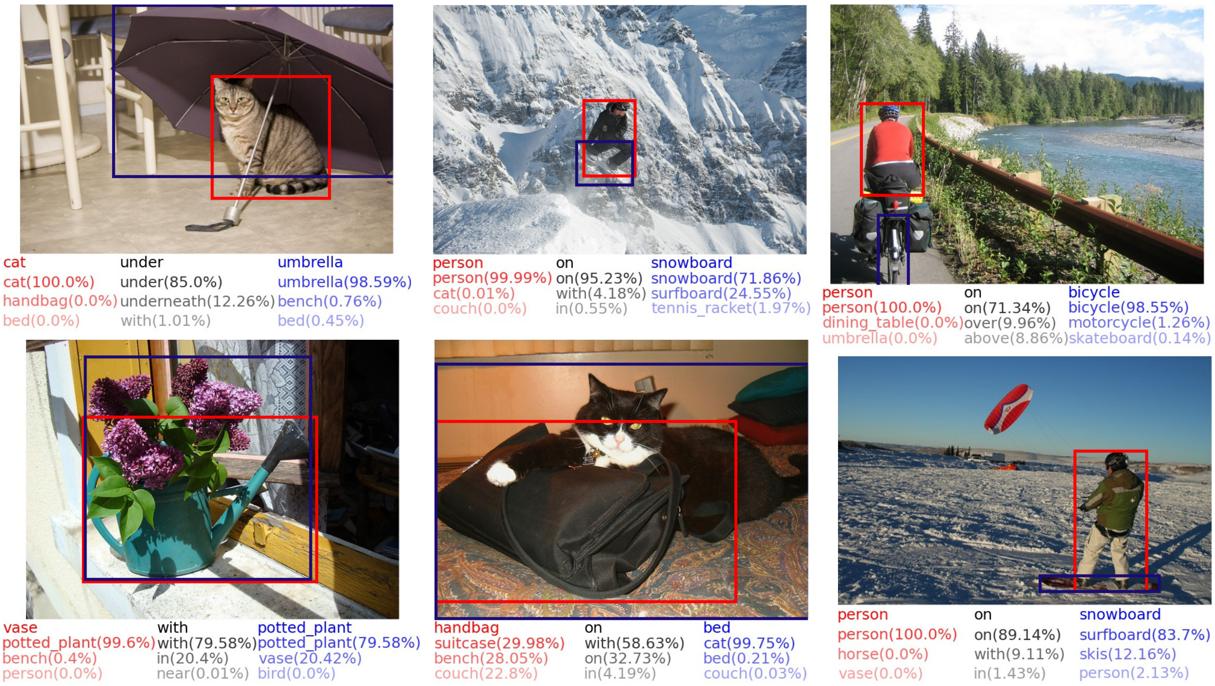

cat under umbrella
cat(100.0%) under(85.0%) umbrella(98.59%)
handbag(0.0%) underneath(12.26%) bench(0.76%)
bed(0.0%) with(1.01%) bed(0.45%)
person on snowboard
person(99.99%) on(95.23%) snowboard(71.86%)
cat(0.01%) with(4.18%) surfboard(24.55%)
couch(0.0%) in(0.55%) tennis_racket(1.97%)
person on bicycle
person(100.0%) on(71.34%) bicycle(98.55%)
dining_table(0.0%) over(9.96%) motorcycle(1.26%)
umbrella(0.0%) above(8.86%) skateboard(0.14%)
vase with potted_plant
potted_plant(99.6%) with(79.58%) potted_plant(79.58%)
bench(0.4%) in(20.4%) vase(20.42%)
person(0.0%) near(0.01%) bird(0.0%)
handbag on bed
suitcase(29.98%) with(58.63%) cat(99.75%)
bench(28.05%) on(32.73%) bed(0.21%)
couch(22.8%) in(4.19%) couch(0.03%)
person on snowboard
person(100.0%) on(89.14%) surfboard(83.7%)
horse(0.0%) with(9.11%) skis(12.16%)
vase(0.0%) in(1.43%) person(2.13%)

Figure 6: More example results for the chain CRF model. The first row below each image shows the ground truth labels for the trajector, preposition and landmark. Subsequent rows show the predicted labels for each, stacked in order of confidence (probabilities for each label in parenthesis).

References

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October. Association for Computational Linguistics.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, December.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC'14*, pages 4585–4592.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisboa, Portugal, September. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, February.