Adding the Third Dimension to Spatial Relation Detection in 2D Images

Brandon Birmingham **Adrian Muscat**
Communications and Computer Engineering
University of Malta
Msida MSD 2080, Malta
adrian.muscat@um.edu.mt

Anja Belz
Computing, Engineering and Mathematics
University of Brighton
Lewes Road, Brighton BN2 4GJ, UK
a.s.belz@brighton.ac.uk

Abstract

Detection of spatial relations between objects in images is currently a popular subject in image description research. A range of different language and geometric object features have been used in this context, but methods have not so far used explicit information about the third dimension (depth), except when manually added to annotations. The lack of such information hampers detection of spatial relations that are inherently 3D. In this paper, we use a fully automatic method for creating a depth map of an image and derive several different object-level depth features from it which we add to an existing feature set to test the effect on spatial relation detection. We show that performance increases are obtained from adding depth features in all scenarios tested.

1 Introduction

Image description aims to produce a summarising description, in structured natural language, of an image (region), typically involving the prioritisation of more important elements and relationships between elements. Work in this area is most commonly motivated in terms of accessibility and data management, and has a range of distinct application tasks. Research in image description and understanding is booming, with relation detection currently a particular focus. The input to spatial relation detection is usually a set of secondary, abstract features derived from region boundaries and labels. A range of different language and geometric features have been used in existing work, but none that explicitly encode information about the third dimension (depth), except via manual annotations (Elliott, 2014). This is an issue for spatial relation detection, because many spatial relations involve three dimensions, some obviously so (e.g. in front of, behind), some less so (beyond, outside, across, etc.). Existing methods in effect try to guess 3D relations from 2D information.

In the experiments in this paper, we use a fully automatic method to generate a depth map from an image, derive different object-level abstract features from the depth values associated with pixels within object bounding boxes, and test the effect of adding such features on the performance of spatial relation detection methods. Below, we start by reviewing related research (Section 2) and describing the existing dataset and associated features we use in our experiments (Section 3). We next describe the depth map generation method we used, and the features we derive from depth maps (Section 4). We then describe the classifier methods we use in experiments (Section 5), and report results from experiments involving different classifier methods and combinations of depth features (Section 6). We conclude with some discussion and a look to the future (Section 7).

2 Related Research

Research on associating text with images goes back at least to the 1960s with early work focusing on object/region labelling (Rosenfeld, 1978). Image description proper starts where a summarising description of the whole image is aimed for. Some approaches measure the similarity of a new image with other images for which descriptions exist, and then use one or more of those descriptions to create a description for the new image (Socher et al., 2014; Karpathy and Fei-Fei, 2015; Ordonez et al., 2011). Our focus here is on methods that create a new description for a given image from scratch. Such methods can be said to involve three main steps: (1) identification of type and, optionally, location of objects and background/scene; (2) detection of attributes, relations and activities involving objects from Step 1; and (3) generation of

Proceedings of The 11th International Natural Language Generation Conference, pages 146–151,
Tilburg, The Netherlands, November 5-8, 2018.

a word string from the outputs from Steps 1 and 2. In Step 2, the focus of this paper, systems determine object attributes (Yatskar et al., 2014; Kulkarni et al., 2011), spatial relationships (Yang et al., 2011; Elliott and Keller, 2013), activities (Yatskar et al., 2014; Elliott and Keller, 2013), etc.

Identifying the spatial relationships between pairs of objects in images is an important part of Step 2, but overlaps into Step 3 if prepositions are selected directly. Methods that produce spatial prepositions sometimes do so as a side-effect of the overall method (Mitchell et al., 2012; Kulkarni et al., 2013); examples of preposition selection as a separate subtask include Elliott and Keller (2013) who base the mapping from features to spatial relations on manually composed rules, and Ramisa et al. (2015) and our own previous work (Muscat and Belz, 2017) where the mapping is learnt automatically. Elliott (2014) manually adds 3rd dimension annotations to images (e.g. whether objects are behind other objects).

There is a sizable literature on spatial relations and spatial language from cognitive and psycholinguistic perspectives, and the remainder of this section briefly surveys a selection of relevant results. Indications are that whether speakers use spatial relations in scene descriptions and referring expressions depends at least in part on individual preference and the context. E.g. when generating referring expressions, some people prefer not to use spatial relations at all (Viethen and Dale, 2008). Furthermore, speakers tend to make more use of spatial relations in domains unknown to them, whereas they use them comparatively less when the domain is known (Viethen and Dale, 2008). Kelleher and Kruiff (2005) categorise spatial relations as combinations of topological vs. projective, and contrastive vs. relative, the latter being dependent on context. Both studies (Viethen and Dale, 2008; Kelleher and Kruijff, 2005) agree that people are generally less likely to use projective spatial relations like *in front of* than topological relations like *on top of*. The former depend on a landmark whereas the latter depend on intersection, overlap and contiguity, which require less cognitive effort to process. For similar reasons, contrastive relations are used more than relative relations (Kelleher and Kruijff, 2005).

The comprehension and choice of spatial prepositions depend on function as well as context (Coventry et al., 2005), e.g. the choice of preposition in *person* ***at*** *a table*, depends on the functional relationship between the trajector object, *person*, and the landmark object, *table*. Dobnik and Kelleher (2014) derive functional semantic knowledge from corpora and use it to explore the dependency of spatial prepositions on functional knowledge.

Regier and Carlson (2001) show that projective spatial terms such as *above* are grounded in attention processes and vector-sum coding of overall direction, formalising these notions in their attentional vector-sum (AVS) model. The model is shown to predict linguistic acceptability judgments for spatial terms, for a variety of spatial configurations. Results indicate that spatial prepositions require more attention on the image compared to detecting an object, and geometric features based on the net vector sum over an area rather than the centre of mass are better predictors.

Kelleher et al. (2011) show that object occlusion degrades the performance of models that are based solely on geometric and functional features e.g. in the case of *in front of*, a projective preposition. Kelleher et al.'s occlusion-enabled regression-based model is shown to outperform Regier and Carlson's AVS model.

3 Data and Features

In the research reported here, we use a subset of the French part of the SpatialVOC2K dataset (Belz et al., 2018), referred to as 'DS-F-Best' below, for consistency with previous publications. Objects in this dataset are annotated with bounding boxes, object labels and spatial relations encoded as sets of prepositions. To create the spatial relation annotations, annotators were asked to (a) choose the single best preposition (free text entry), as well as (b) select all possible prepositions from a list of candidate spatial prepositions, such that the preposition(s) accurately describe(s) the spatial relationship between the given pair of objects.

In the experiments below, we are interested in studying the effect depth features have on recalling individual prepositions (especially the ones that have previously proven difficult to predict) in addition to the overall system-level recall. We therefore use the single *best* preposition for each object pair only, when training the single label classifiers.

In research involving this and similar datasets, sets of language and geometric features are normally computed from bounding boxes and object labels. Typical language features are label en-

coders (one hot vectors) and word2vec (Mikolov et al., 2013) vectors. Examples of geometric features are area of object bounding box normalised by combined area size for both objects, area of overlap between the two bounding boxes normalised by combined area, and Euclidean distance between two bounding boxes. Some of the feature functions are unary and others are binary. For the initial feature set in this paper, we used the union of geometric features from two previous lines of work, our own (Muscat and Belz, 2017) and Ramisa et al. (2015). This yielded a set of 18 geometric features, and although some of these are correlated, we left it to the classifier models to discriminate among the more useful ones. There are no 3D features in this initial set of features, although some features are intended as proxy features for depth, e.g. bounding box overlap.

4 Computing Depth Features

4.1 MonoDepth Features

We use *monoDepth*[1] (Godard et al., 2017), a convolutional neural network method trained on stereo image pairs which maps single images to depth maps where each pixel has a value assigned to it that represents the estimated distance from the viewer. More specifically we used the monodepth-cityscapes model, trained on the Cityscapes dataset (Cordts et al., 2016). Figure 1 shows an image from our dataset alongside the depth map generated for it by the monodepth-cityscapes model. The more towards the dark blue end of the colour spectrum an area is, the further away it is from the viewer, and the more towards the bright yellow end, the nearer. The model produces an impressively accurate rendering of the depths of the two trees, car, person, and road (not all depth maps are as good).

Once we have the depth map for a given image, we obtain depth values for the pixel grids inside the bounding boxes (BBs) of the pair of objects under consideration. We then compute the following object-level features for each BB:

- Average depth (AVG): simply the average depth value within each object BB.
- Radially weighted average (RWA) depth: starting from the central pixel(s), assign a weight to each pixel that is in inverse proportion to its distance from the centre, then compute the weighted average.

Looking at the example in Figure 1, AVG is much lower in the red person BB than in the blue car BB, making 'person in front of car' a possibility. RWA is also less for the person BB, but the difference is less pronounced than would be the case if all of the car was further way than the person, thus making 'person next to car' an alternative possibility.

4.2 Human-estimated Depth Feature

We obtained human estimates of BB-level depth for 1,554 images and 3,642 objects as follows. Participants were shown an image with objects surrounded by BBs. Their task was to assign a number out of 100 to each bounding box, indicating the average depth of (just) the object inside the BB, where 100 is the maximum distance. The annotators were trained and mentored for some time before starting annotations proper. Three participants in total contributed to the annotations. Depth values were then normalised to range from 0 to 1 for each image.

We computed Pearson's correlation coefficients between the human estimated object depths and the corresponding AVG and RWA figures. Pearson's r between human and AVG depth values was 0.535 $(p < 0.0001)$, and between human and RWA it was 0.523 $(p < 0.0001)$. The correlation between AVG and RWA was $0.995 (p < 0.0001)$. We also converted the three sets of depth estimates to categorical values (foreground, background, neutral) and computed percentage agreement with human-estimated depth on these, which was 60.8% for AVG and 60.3% for RWA.

5 Methods

Using combinations of features from Section 3 and 4, we separately trained models of the six types below.[2] Where relevant, hyperparameters for the models were obtained by splitting the development data into separate training and validation sets, which were then recombined for training the final models and testing on a held-out test set. All models output the probability vector for the prepositions, from which results are calculated.

Naive Bayes (NB) models assume that each feature is conditionally independent of every other feature given the output class (preposition in our case). We use a prior computed from the output

[1] https://github.com/mrharicot/monodepth

[2] Using scikit-learn: http://scikit-learn.org

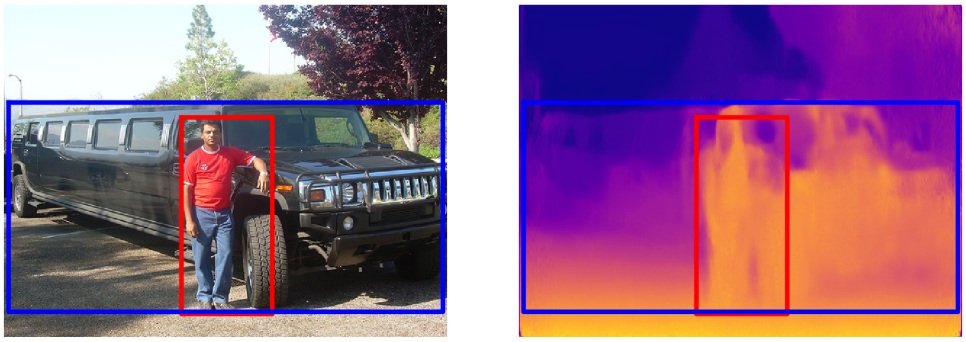

Figure 1: Example SpatialVOC2K image and depth map generated by monoDepth.

labels, and base the likelihood on the geometric features.

Decision Tree (DT): Decisions are based on conjunctions of features. Values for the maximum tree depth [2, 20] are determined by hyperparameter optimisation (HPO).

Logistic Regression (LR): A linear classifier which models the SR probabilities with a logistic function. The value for the inverse of regularisation constant [0.1, 100.0] is determined by HPO. The regularisation is L1-norm, tolerance is 0.001 and one-versus-rest multi-class classification.

Support Vector Machine (SVM): A binary classifier solving the multiclass case via (here) one-versus-one classification. The RBF kernel parameters, C [0.1, 100.0] and gamma [0.001, 1.0] are determined by HPO.

Random Forests (RF): A meta-estimator comprising multiple decision-tree classifiers fitted to sub-samples of the data, using averaging to improve predictive accuracy and to control overfitting. The number of estimators [10, 150], maximum features [1, 156], maximum tree depth [2, 20], are determined by HPO.

6 Experiments and Results

We carried out experiments for all ML methods above, and for the following feature combinations: (i) the 18 geometrical features ('G' in results tables) from Section 3, (ii) the language features derived from the object labels ('L' in the tables), (iii) average depth ('avg' in tables), (iv) RWA ('rwa' in tables) and (V) human-estimated depth ('man' in tables). For each of (iii), (iv) and (v) we considered depth of object 1 ('d1' in tables), depth of object 2 ('d2' in tables), and the difference between the latter two depths ('dd' in tables).

Table 1 shows system-level weighted average recall results. Depth features improved the weighted average recall results across the board. The highest increase is 8.9% when added to geometric features, and 8.3% when added to both language and geometric features. AVG and RWA features perform equally well, and less well than the human-estimated depths. Out of the three depth features, the difference in depth ($dd = d1 - d2$) has the most pronounced positive effect on scores individually; however, the overall highest scores are obtained when all three (d1, d2 and dd). Out of the different classifier modesl, the RF model resulted in the highest scores followed by LR, SVM, DT and NB. However, the NB model registered the highest increase in scores resulting from depth features: 12.5% when added to geometric features.

Features	RF	DT	LR	SVM	NB
G	0.45	0.36	0.4	0.38	0.24
+avg:d1,d2	0.45	0.36	0.4	0.39	0.25
+avg:dd	0.45	0.36	0.4	0.37	0.27
+avg:d1,d2,dd	**0.46**	0.36	0.39	0.37	0.27
+rwa:d1,d2	0.45	0.36	0.4	0.37	0.24
+rwa:dd	0.45	0.35	0.4	0.38	0.27
+rwa:d1,d2,dd	**0.46**	0.35	0.4	0.37	0.26
+man:d1,d2	0.47	0.36	0.4	0.4	0.24
+man:dd	0.47	0.39	0.41	0.4	0.27
+man:d1,d2,dd	**0.49**	0.39	0.4	0.4	0.27
L,G	0.48	0.4	0.46	0.43	0.26
+avg:d1,d2	0.5	0.4	0.46	0.46	0.27
+avg:dd	0.49	0.4	0.46	0.46	0.26
+avg:d1,d2,dd	0.5	0.4	0.46	0.45	0.27
+rwa:d1,d2	0.48	0.4	0.46	0.44	0.27
+rwa:dd	0.48	0.4	0.47	0.44	0.26
+rwa:d1,d2,dd	0.47	0.4	0.46	0.45	0.27
+man:d1,d2	0.49	0.4	0.48	0.46	0.27
+man:dd	**0.52**	0.42	0.47	0.44	0.26
+man:d1,d2,dd	0.51	0.42	0.48	0.44	0.27

Table 1: SpatialVOC2K: Weighted Average Recall for all feature combinations (for explanation of abbreviations, see in text).

Features	a_cote_de	a_l'exterieur_de	au_dessus_de	au_niveau_de	autour_de	contre	dans	derriere	devant	en_face_de	loin_de	pres_de	sous	sur	wt_m	mean
G, avg:d1,d2	-4	-	-20	+8	0	-27	0	+10	+8	0	0	-24	-8	+5	0	-2
G, avg:dd	-4	-	-20	0	0	-27	0	+24	-5	0	+4	-24	-5	0	0	-2
G, avg:d1,d2,dd	-4	-	0	+8	0	0	0	+19	0	0	-6	0	-5	+5	+2	0
G, rwa:d1,d2	-4	-	-20	+8	0	0	0	+19	0	-16	-6	-36	-5	+8	0	-2
G, rwa:dd	0	-	+20	+8	0	+27	0	-5	-5	0	+4	-16	-8	+2	0	0
G, rwa:d1,d2,dd	0	-	-20	-8	0	+27	-25	+24	-5	0	-9	+20	-2	+5	+2	-2
G, man:d1,d2	-4	-	0	0	0	+27	0	+33	0	0	0	0	-2	-2	+4	+7
G, man:dd	0	-	0	+8	0	0	-25	+10	+32	-20	0	+28	0	+2	+4	0
G, man:d1,d2,dd	-4	-	0	+15	0	-27	0	+24	+12	0	+4	+4	+2	+8	+9	+9
G, L	+4	-	0	-15	+33	+73	+24	+5	+18	+20	-9	0	0	+11	+7	+7
G,L,avg:d1,d2	0	-	-20	+18	-25	0	0	+5	0	+7	+10	+20	-2	0	+4	+7
G,L,avg:dd	-4	-	-20	+18	-25	+15	0	+23	-15	0	0	0	0	0	+2	0
G,L,avg:d1,d2,dd	-4	-	-20	+27	-25	-27	0	+18	-9	0	+10	+12	0	+4	+4	+2
G,L,rwa:d1,d2	0	-	-20	+9	-25	+15	0	+5	+4	+7	+4	-16	-2	-2	0	0
G,L,rwa:dd	-4	-	-20	+27	-25	-15	0	+14	-4	-10	+10	-16	0	0	0	0
G,L,rwa:d1,d2,dd	0	-	-20	0	-25	-15	0	+9	-15	-17	+4	-8	0	-2	-2	0
G,L,man:d1,d2	0	-	-20	+18	-25	-15	0	+14	-9	+20	+14	-8	-2	0	+2	+2
G,L,man:dd	+4	-	-20	+55	0	0	0	+14	+9	0	+14	+12	0	0	+8	+7
G,L,man:d1,d2,dd	-4	-	-20	+36	0	-27	0	+23	+4	0	+27	0	+2	+2	+6	+9

Table 2: SpatialVOC2K: Percentage increase in recall per preposition for the RF model. Figures in top half relative to geometric features; lower half relative to both geometric and language features.

This could indicate that the other models are learning more about depth from the other features.

Table 2 shows per-preposition weighted average recall results. In this set of results we examine the effect of adding depth information on individual prepositions, looking at which combinations of features increase or decrease the recall per preposition. The table is split into two halves. The top half shows changes from adding depth features to (just) the geometric features (G), while the bottom half shows changes from adding depth features to the union of geometric and language features (G,L). Some prepositions fare better with depth information: *au niveau de* ("at the level of"), *derriere* ("behind"), *devant* ("in front of"), *sur* ("on"). Results for others worsen: *à côté de* ("next to"), *en face de* ("facing"), *sous* ("under"). For some, the results are inconclusive (*contre* ("against"), *dans* ("in"), *loin de* ("far from"), *près de* ("near")), while others are not affected (*au dessus de* ("above"), *autour de* ("around")).

The row labelled 'G,L' shows the effect of just adding language features to the geometric set. Some prepositions (most notably *autour de, contre* and *dans*) benefit substantially from language features while others benefit more from depth features. Some (*au niveau de, oin de*) fare worse when language features are added. The biggest improvement when depth information is added to geometric features is 33% for *derriere* ("behind"); the highest when depth is added to both geometrical and language is 55%, for *au niveau de* ("at the level of, at equal distance from the viewer").

Getting improvements for clearly 3D prepositions such as *derriere, devant* and *au niveau de* is as expected, but there are clear improvements for other prepositions too.

7 Conclusion

We have reported the first results for using object-level depth features computed from depth maps automatically generated for a given image with monoDepth as additional features in spatial relation prediction. We have shown that performance increases when depth features are added in all scenarios tested. However, automatically computed depth is still some way off manual toplines which resulted in bigger improvements.

References

A. Belz, A. Muscat, P. Anguill, M. Sow, G. Vincent, and Y. Zinessabah. 2018. Spatialvoc2k: A multilingual dataset of images with annotations and features

for spatial relations between objects. In *Proceedings of INLG'18*.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In *Spatial Cognition IV. Reasoning, Action, Interaction*, pages 98–110, Berlin, Heidelberg. Springer Berlin Heidelberg.

Simon Dobnik and John Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third Workshop on Vision and Language*, pages 33–37. Dublin City University and the Association for Computational Linguistics.

D. Elliott and F. Keller. 2013. Image description using visual dependency representations. In *Proc. 18th Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1292–1302, Seattle.

Desmond Elliott. 2014. *A Structured Representation of Images for Language Generation and Image Retrieval*. Ph.D. thesis, University of Edinburgh.

Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*.

A. Karpathy and L. Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3128–3137, Boston.

John Kelleher and Geert-Jan Kruijff. 2005. A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

John D. Kelleher, Robert J. Ross, Colm Sloan, and Brian Mac Namee. 2011. The effect of occlusion on the semantics of projective spatial terms: a case study in grounding language in perception. *Cognitive Processing*, 12(1):95–108.

G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1601–1608, Colorado Springs.

G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, Lake Tahoe, Nevada. Curran Associates Inc.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé, III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proc. of the 13th Conf. of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France.

A. Muscat and A. Belz. 2017. Learning to generate descriptions of visual data anchored in spatial relations. *IEEE Computational Intelligence Magazine*, 12(3):29–42.

V. Ordonez, G. Kulkarni, and T. L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, Granada, Spain.

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 *Proc. 20th Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 214–220, Lisbon, Portugal.

Terry Regier and Laura A. Carlson. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology General*, 130(2):273–298.

A. Rosenfeld. 1978. Iterative methods in image analysis. *Pattern Recognition*, 10(3):181–187.

R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:27–218.

J. Viethen and R. Dale. 2008. The use of spatial relations in referring expression generation. In *Proc. 5th Int. Natural Language Generation Conf. (INLG)*, pages 59–67, Salt Fork, Ohio.

Y. Yang, C. L. Teo, H Daumé III, and Y. Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proc. 16th Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 444–454, Edinburg, Scotland.

M. Yatskar, L. Vanderwende, and L. Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Proc. 3rd Joint Conference on Lexical and Computational Semantics*, pages 110–120, Dublin, Ireland.