Learning Visually Grounded Common Sense Spatial Knowledge for Implicit Spatial Language*

Guillem Collell and **Marie-Francine Moens**
Computer Science Department
KU Leuven
gcollell@kuleuven.be; sien.moens@cs.kuleuven.be

1 Motivation

Spatial understanding is crucial for any agent that navigates in a physical world. Computational and cognitive frameworks often model spatial representations as *spatial templates* or regions of acceptability for two objects under an *explicit* spatial preposition such as "left" or "below" (Logan and Sadler 1996). Contrary to previous work that define spatial templates for *explicit* spatial language only (Malinowski and Fritz 2014; Moratz and Tenbrink 2006), we extend such concept to *implicit* spatial language, i.e., those relationships (usually actions) that do *not* explicitly define the relative location of the two objects (e.g., "dog *under* table") but only implicitly (e.g., "girl *riding* horse"). Unlike *explicit* relationships, predicting spatial arrangements from *implicit* spatial language requires spatial common sense knowledge about the objects and actions. Furthermore, prior work that leverage common sense spatial knowledge to solve tasks such as visual paraphrasing (Lin and Parikh 2015) or object labeling (Shiang et al. 2017) do not aim to predict (unseen) spatial configurations.

Here, we propose the task of predicting the relative spatial locations of two objects given a textual input of the form (Subject, Relationship, Object). We report on initial experiments with a simple neural network model with distance-based supervision learned in annotated images that obtains promising performance. Crucially, we show that the model can reliably predict templates of **unseen combinations**, e.g., predicting (man, riding, elephant) without having seen such scene before. Furthermore, by leveraging word embeddings of objects and relationships, the model can correctly predict spatial templates for **unseen words**. E.g., without having ever seen "boots" before but only "sandals", the model predicts correctly the template of (person, wearing, boots) by inferring that, since "boots" are similar to "sandals", they must be worn at the same position of the "person"'s body. Hence, the model is able to leverage the learned common sense spatial knowledge to generalize to unseen objects.

*The reader may refer to a full paper (Collell, Van Gool, and Moens 2018) that resulted from the preliminary studies presented in this abstract.

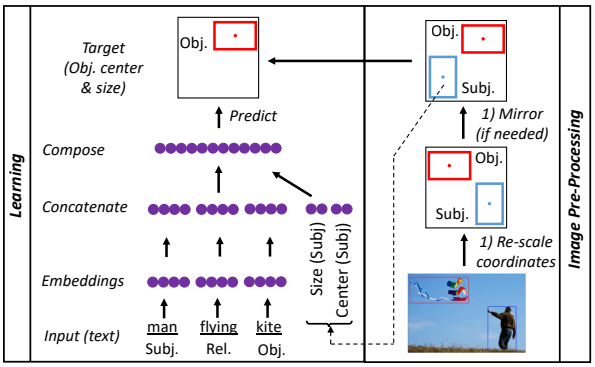

Figure 1: Overview of our model and setting.

2 Proposed task and model

2.1 Proposed task

We propose the task of predicting the 2D relative spatial arrangement of two objects under a relationship given a structured text input of the form (Subject, Relationship, Object)—abbreviated as (S, R, O). More precisely, the model predicts the Object's box *center* and box *size* (**output**) given the structured text **input** (S, R, O) plus the *center* and *size* of the Subject's box (Fig. 1).

2.2 Proposed model

We employ a feed forward network with embeddings (Fig. 1). The **embedding layer** maps the input words (S,R,O) to their d-dimensional representations. The embeddings are then concatenated with the Subject's box *center* and *size*. This vector is then fed into a fully connected layer to **compose** S, R, O into a joint representation. model predictions (Object's *center* and *size*) are evaluated against ground truth with a mean squared error (MSE) loss.

3 Experimental setup

Data. We use the Visual Genome (Krishna et al. 2017) dataset, which has ∼108K images containing ∼1,5M human-annotated (S, R, O) instances with corresponding object boxes. We filter out all instances with *explicit* spatial prepositions, yielding ∼378K *implicit* (S, R, O) instances.

		MSE	R^2	acc_y	$F1_y$	r_x	r_y
Implicit	*EMB*	**0.008**	0.705	0.756	0.755	0.894	0.834
	RND	**0.008**	0.691	0.750	0.750	0.891	0.826
	1H	**0.008**	**0.717**	**0.762**	**0.762**	**0.896**	**0.842**
	ctrl	0.054	-1.000	0.522	0.521	0.000	-0.001
Explicit	*EMB*	0.013	0.586	0.768	0.770	0.811	0.823
	RND	0.013	0.580	0.767	0.769	0.808	0.815
	1H	**0.012**	**0.604**	**0.778**	**0.780**	**0.815**	**0.828**
	ctrl	0.060	-1.000	0.633	0.630	0.000	0.000

Table 1: Results on **implicit** and **explicit** relations.

Evaluation sets. We evaluate performance in the following subsets of Visual Genome. **(i) Raw** set: Simply the unfiltered instances. **(ii) Unseen words**: We randomly pick 25 objects (e.g., "woman", "apple", etc.) among the 100 most frequent ones and leave out from the training data all the instances (∼130K) containing any of these words. This set is used for testing. **(iii) Unseen combinations**: We randomly pick 100 combinations (S, R, O) among the 1,000 most frequent *implicit* ones and leave them out for training. We finally consider the *explicit* version of the **Raw** set. Reported results are always on unseen *instances*—yet the *combinations* (S, R, O) may have been seen during training (e.g., in different images).

Data pre-processing. Coordinates are normalized by image width and height. Since right/left depends only on the camera viewpoint, we get rid of this arbitrariness by *mirroring* the image when the Object is on the left of the Subject.

Evaluation metrics. We use standard ***regression*** metrics: **(i) Mean Squared Error (MSE)** between predicted and true Object center and size. **(ii) Coefficient of Determination (R^2)** of model predictions and ground truth. **(iii) Pearson Correlation (r)** between predicted and true x-component of the Object center, and similarly for the y-component. We also consider the ***classification*** of above/below relative locations of the Object w.r.t. the Subject. We report (macro averaged) **F1 ($F1_y$)** and **accuracy (acc_y)**.

4 Results

We test the following model variations. *EMB* denotes a model that uses pre-trained word embeddings[1], *RND* a model with random normal embeddings, *1H* employs one-hot embeddings and *ctrl* outputs random normal predictions. Overall, the preliminary results outlined below look promising.

4.1 Quantitative results

Evaluation with raw data. Table 1 shows that all methods perform well in the *Raw* data. Remarkably, we see that relative locations can be predicted from *implicit* spatial language at least as accurately as from *explicit* spatial language.

Unseen combinations. All models perform well on unseen combinations (table not shown), remarkably closely to their performance with seen combinations.

Unseen Words. Contrarily, large differences in performance are observed with unseen words (table not shown) where the model that uses embeddings (*EMB*) performs significantly better than the rest.

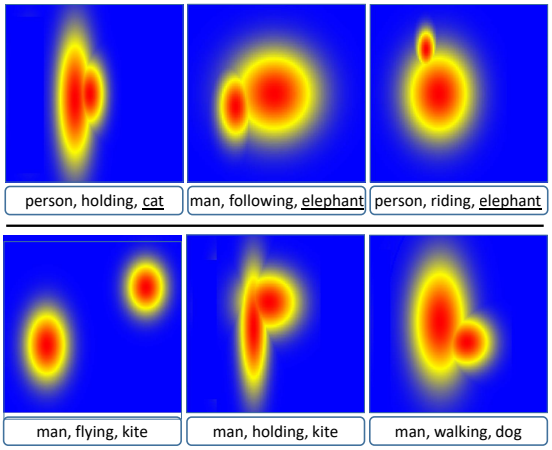

Figure 2: Predictions by the model that leverages word embeddings (*EMB*). **Top**: Predictions in unseen words (underlined). **Bottom:** Predictions in unseen *triplets*.

4.2 Qualitative evaluation (spatial templates)

Heat maps in Fig. 2 show regions of predicted high (red) and low (blue) probability. The "heat" of the objects is assumed to be normally distributed with μ equal to the object's center and σ to the object's size. The *EMB* model is able to infer both, relative locations and sizes, e.g., predicting correctly the size of a "cat" relative to a "person" even though the model has never seen a "cat" before. Notably, the model learns to compose the triplet as a whole, distinguishing, e.g., (man, flying, kite) from (man, holding, kite).

Acknowledgments

This work has been supported by the CHIST-ERA EU project MUSTER.[2]

References

Collell, G.; Van Gool, L.; and Moens, M.-F. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. AAAI.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Lin, X., and Parikh, D. 2015. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *CVPR*, 2984–2993.

[1]We use 300-d GloVe embeddings (Pennington, Socher, and Manning 2014) http://nlp.stanford.edu/projects/glove.

[2]http://www.chistera.eu/projects/muster

Logan, G. D., and Sadler, D. D. 1996. A computational analysis of the apprehension of spatial relations.

Malinowski, M., and Fritz, M. 2014. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv preprint arXiv:1411.5190*.

Moratz, R., and Tenbrink, T. 2006. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition and computation* 6(1):63–107.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, 1532–1543.

Shiang, S.-R.; Rosenthal, S.; Gershman, A.; Carbonell, J.; and Oh, J. 2017. Vision-language fusion for object recognition. AAAI.