Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2368–2372
Marseille, 11–16 May 2020

Understanding Spatial Relations through Multiple Modalities

Soham Dan, Hangfeng He, Dan Roth
University of Pennsylvania
{sohamdan, hangfeng, danroth}@seas.upenn.edu

Abstract

Recognizing spatial relations and reasoning about them is essential in multiple applications including navigation, direction giving and human-computer interaction in general. Spatial relations between objects can either be *explicit* – expressed as spatial prepositions, or *implicit* – expressed by spatial verbs such as *moving*, *walking*, *shifting*, etc. Both these, but implicit relations in particular, require significant common sense understanding. In this paper, we introduce the task of inferring *implicit* and *explicit* spatial relations between two entities in an image. We design a model that uses both textual and visual information to predict the spatial relations, making use of both positional and size information of objects and image embeddings. We contrast our spatial model with powerful language models and show how our modeling complements the power of these, improving prediction accuracy and coverage and facilitates dealing with unseen subjects, objects and relations.

Keywords: Knowledge Discovery/Representation

Figure 1: woman [?] bed. Language models adopt the choice seen most commonly, i.e., sleeping on, but we propose an image-specific model.

1. Introduction

Humans are able to do common sense reasoning across a variety of modalities – textual, visual and for a variety of tasks – reasoning, locating, navigation. Several such tasks require spatial knowledge understanding and reasoning (Kordjamshidi et al., 2010), (Kordjamshidi et al., 2011), (Johnson et al., 2017), (Wang et al., 2017), (Baldridge et al., 2018). Although, there has been several recent works on common-sense reasoning (Speer and Havasi, 2012), (Emami et al., 2018), (Sap et al., 2018), (Storks et al., 2019), progress on spatial understanding and common-sense is rather limited. Prior work has either not been spatially focused (Lin and Parikh, 2016), (Xu et al., 2015),(Antol et al., 2015) or very restrictive in the class of spatial relations they handle (Xu et al., 2017), (Kordjamshidi et al., 2017). Recently, (Collell et al., 2018) presented the task of predicting an object's location and size in an image given the subject's bounding box and the spatial relation between them.

In this paper, we address the problem of understanding spatial relations. We specifically want to infer the spatial relationship between two entities given an image involving them. Spatial relations can either be – *explicit* (spatial prepositions such as *on*, *above*, *under*) or *implicit* (intrinsic spatial concepts associated with actions – *sleeping*, *sitting*, *flying*). We take as input an image and the bounding boxes of the entities which are spatially related, the word and image embeddings of the two entities, and we want to predict the spatial relation between them (eg: *sleeping*, *standing*, *sitting-on* in Figure 1. We compare this with powerful language models like BERT (Devlin et al., 2018) which have been trained on very large text corpora in a variety of contexts. Although BERT is not conditioned on the image, it can still provide a good list of candidate relations between the two entities using only the structured text information and it can provide new relations unseen in the specific dataset we train on. We show that these complementary attributes – being able to use the image-specific information by a task-specific model and being able to predict a wide range of relations – by BERT, can together give better performance than either approach alone, especially in low-resource and generalized settings (Collell et al., 2018). For the *woman,?,bed* example, if we were to rely on a language model alone, the prediction would be *laying on* or *sleeping on* almost all the time. This is because using just the textual modality makes it blind to the image specific cues. Using both the textual and visual information leads to a more robust model conditioned on both the image and the subject and object text description.

Our contributions are four-fold: **(1)** New task definition of *explicit* and *implicit* spatial relation prediction for two entities in an image. We explore another dimension of commonsense understanding of spatial relations with visual and language information. **(2)** Usage of **Spatial BERT**: combination of BERT and a spatial model. We conduct thorough experiments to show the role of the image, position and language information for the task under different settings-varying the number of training examples, the type of the spatial relations and the type of BERT.[1] **(3)** As a byproduct, we propose a re-scoring technique for evaluating this model combination. **(4)** We show that **Spatial BERT** is

[1] https://github.com/sdan2/Multimodal-Spatial

able to predict in the generalized setting – for unseen subjects, objects or relations.

2. Model Details

2.1. The Basics

This task is to predict the spatial relation R given the subject S and the object O. For the subject S, we have the corresponding text information T_s, position information P_s, and image information I_s. Similarly, we have the corresponding text information T_o, position information P_o, and image information I_o for the object O.

2.2. The Spatial Model

Feed Forward Network (FF). For the text information, we use average (for multi-word subjects and objects) glove embeddings (Pennington et al., 2014) of the words in the text to represent it. For simplicity, we use w_s and w_o to represent the average glove embbeddings for the T_s and T_o. As for the position P, it contains 4 float values denoting the x and y coordinate of the subject(object) center and the half-width and half-height of the bounding box of the subject(object). We then pass w_s, w_o, P_s, P_o through a feed forward network with 128 neurons and take a softmax of the predictions of all possible relations:

$$h = [\sigma(W_s[w_s; P_s] + b_s); \sigma(W_o[w_o; P_o] + b_o)]$$

$$\hat{p} = softmax(W_h h + b_h)$$

where σ is the activation function ReLU.

Feed Forward Network + Image Embeddings (FF+I). In addition to the text information and position information of the subject and object, we use pre-trained visual embeddings to represent the image information I_s and I_o. For simplicity, we use i_s and i_o to represent the visual embeddings for the subject and the object. The visual embeddings of the words are provided by (Collell and Moens, 2018) from a VGG128 network (Chatfield et al., 2014) pre-trained on Imagenet and fine-tuned on Visual Genome (Krishna et al., 2016). The hidden representation in the FF + Image Embeddings are as following:

$$h = [\sigma(W_s[w_s; P_s; i_s] + b_s); \sigma(W_o[w_o; P_o; i_o] + b_o)]$$

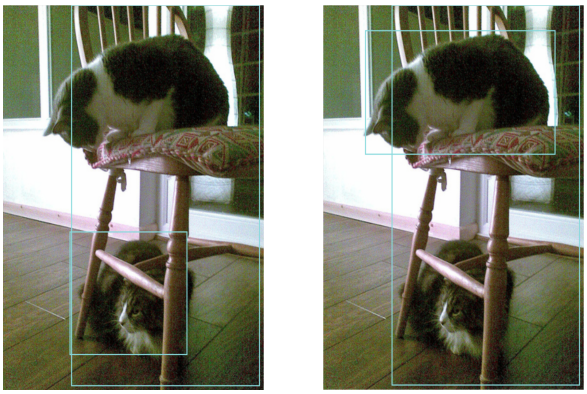
2.3. The Language Model

BERT. We use BERT (Devlin et al., 2018) as language model to predict the most likely spatial relation by masking the relation and providing "subject [MASK] object" as input and running the beam search to obtain top 20 predictions.

Fine-tuned BERT (f-BERT). We fine-tune BERT for the Visual Genome datatset by collecting all the "subject relation object" texts from the training data.

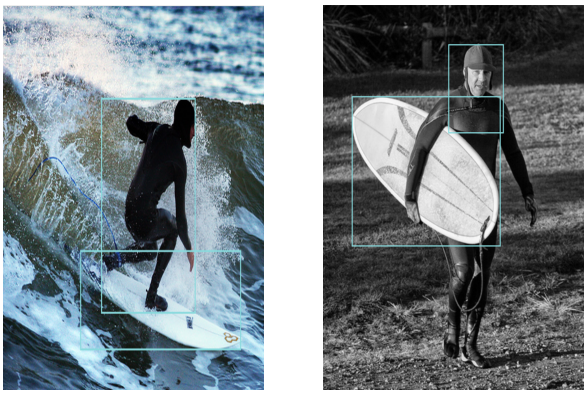
2.4. Spatial BERT

We combine the prediction of our best model (FF+I) with normal BERT and fine-tuned BERT to get two combined models. We essentially re-rank the predictions from the two models. Assume $\hat{p}_{bert}$ and $\hat{p}_{ff}$ are the predicted probability distribution from the BERT and FF+I models, the predicted probability of Spatial BERT is: $\hat{p} = \hat{p}_{ff} + \lambda \hat{p}_{bert}$, where λ is a non-negative float to adjust the weight of the BERT predictions.



(a) cat UNDER chair (b) cat ON chair

Figure 2: Examples of explicit spatial relations from Visual Genome of $(cat, ?, chair)$



(a) man RIDING surfboard (b) man CARRYING surfboard

Figure 3: Examples of implicit spatial relations from Visual Genome of $(man, ?, surfboard)$

3. Experiments

3.1. Dataset

We use the Visual Genome dataset (Krishna et al., 2016). We work on the (subject, relation, object) triples where the subject and object is accompanied by the bounding box information[2]. The dataset is partitioned into two categories: *explicit* and *implicit* based on the spatial relation in a triple and the experiments are performed separately for each category. There are 333321 implicit and 682374 explicit triples. In Figure 2 we see examples from the dataset for the $explicit$ relations between cat as $subject$ and $chair$ as $object$. In Figure 3 we see examples from the dataset for the $implicit$ relations between man as $subject$ and $surfboard$ as $object$.

For the *explicit* relations, *on* is the majority relation with frequency 425308 and **the majority baseline for *explicit* relation prediction is 62%**. Similarly, for the *implicit* relations, *has* is the majority relation with a frequency 167780

[2]Note each image is scaled to 1 so that the sizes are comparable. See (Collell et al., 2018) for more information on the data-preprocessing step.

Figure 4: cat BENEATH chair. Even if we have never seen *beneath* as a relation during the training phase but have seen cat UNDER chair (Figure 2), using the BERT predictions and re-scoring the choices using the GloVe embedding similarity, we want to be able to predict unseen relations at test time.

and **the majority baseline for *implicit* relation prediction is 50%**.

3.2. GloVe based Re-Scoring Metric

One of the principal benefits of using language models is that they can predict new relations never seen during the training of the spatial model (see Figure 4). In several cases (especially the low training data regimes) where BERT is used in combination with the spatial model, BERT may predict spatial relations which are unseen in training and we need an effective strategy to re-score the spatial model predictions based on these unseen predictions. For example, say, BERT predicts *atop* (not seen during training) as the relation between *book* and *table* for an image whose gold triple is (*book*, *on*, *table*). In such situations we develop a re-scoring metric to distribute the score BERT assigns to unseen relations (*atop*) among the seen relations (*on*, *above*, *over*) which are related to it. This relatedness is measured by the cosine similarity between the unseen word vector and the word vectors for relations present in the dataset. Thus, $score(on) = score(on) + sim(on, atop) * score(atop)$.

3.3. Experimental Settings

We present the experiment results separately for the explicit and implicit relations. For each type of relation, we vary the percentage of data used for training as 1%, 10%, 50%, 75%, 100%. We try two variations of BERT – the normal pre-trained BERT and f-BERT: BERT fine-tuned on the *implicit* and *explicit* dataset respectively. We try variety of λ (from 0.01 to 1) to combine the BERT scores and the spatial model scores and we report the best result across different values of λ (we exclude 0 or ∞ (large positive values) which are already reflected in only-BERT and only the spatial model results). In all the settings the train-development-test split is set as $75 - 15 - 15\%$.

3.4. Results

We see that the best performance in several of the settings is achieved by Spatial BERT. Although the gains may look small, the number of data-points is huge and thus, the improvement is significant in terms of absolute counts. For smaller percentages of training data, the spatial models perform poorly but later on beats BERT. Also, notice BERT performs significantly better for the explicit spatial relations than the implicit spatial relations possibly because the set of implicit relations is much larger and the task of predicting them requires more image understanding and common-sense compared to the explicit spatial relations. BERT and f-BERT performances do not change across different data percentages because they are both just used for inference and the amount of training data does not affect them.

4. Unseen Subject, Object or Relation

It is greatly desirable that models learn to generalize to unseen contexts and is necessary for true spatial understanding of the relations.

4.1. The Settings

If a spatial relation was seen during training – (*man*, *riding*, *horse*) and the supporting image, the model should be able to infer that the *implicit* (in this example) spatial relation for a new image depicting (*lady*, *riding*, *elephant*) should be *riding*, even if it has never seen the subject (lady) or object (elephant) before. We show two illustrative examples from Visual Genome that we want to handle for the unseen subject (Figure 5) and unseen object (Figure 6) settings respectively. The pre-trained embeddings of the unseen subject(object) is similar to the embeddings of similar seen subject(object) and this (along with the positional information) should help the model identify that the relations are similar. To systematically test this capability, we perform experiments for the *implicit* and *explicit* relations in Table 2. For each dataset type we experiment with three settings – **unseen subject, unseen object and unseen relation**.

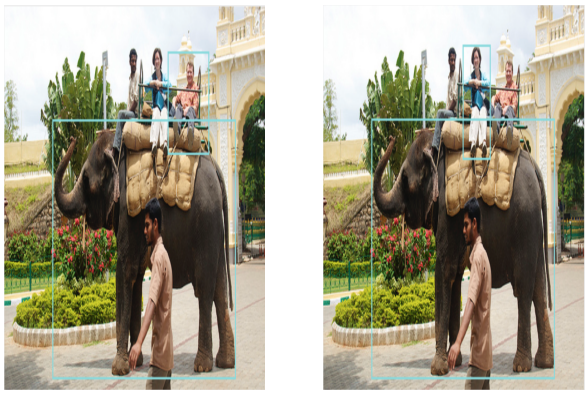
The unseen subject(object) setting is relatively easier compared to the unseen relation setting. For subject, we test if we can correctly predict the *flying* relation in (*kid*, *flying*, *kite*) even if we have never seen (*kid*, *flying*, X) during training. For objects, we test if we can correctly predict the *riding* relation in (*man*, *riding*, *elephant*) even if we have never seen (X, *riding*, *elephant*) during training. Here, X denotes any object or subject, respectively. We first tabulate all the (subject, relation), (object,relation) pairs and we split this list into the test set pairs(15%) and the training and development set pairs(85%). We then form the test set (train, development) by collecting all datapoints whose (subject,relation), (object,relation) is in the test(train or development) set pairs. Thus, the test set has (subject,relation), (object,relation) which are not seen during training. We only use the normal pre-trained BERT for the generalized experiments since fine-tuning is not very natural for these settings.

Spatial Relation	Explicit					Implicit				
% of Training Data	1%	10%	50%	75%	100%	1%	10%	50%	75%	100%
BERT	38.56	38.56	38.56	38.56	38.56	6.9	6.9	6.9	6.9	6.9
f-BERT	73.03	73.03	73.03	73.03	73.03	**77.6**	**77.6**	77.6	77.6	77.6
FF	0.8	72.1	74	74.4	74.7	0	75.0	78.8	79.0	79.5
FF+I	0.71	72.7	74.1	74.5	74.72	0.01	75.79	78.86	79.3	79.5
BERT+FF+I	36.4	72.88	**74.7**	**74.9**	**75.2**	6.35	75.7	79	79.3	79.5
f-BERT+FF+I	**73.06**	**73.06**	74.2	74.52	74.74	**77.6**	77.5	**79.3**	**79.7**	**79.9**

Table 1: Comparison of performance (in percentage accuracy) of different models for explicit and implicit spatial relations, normal and f-BERT and combinations with the best spatial model for varying portions of training data.

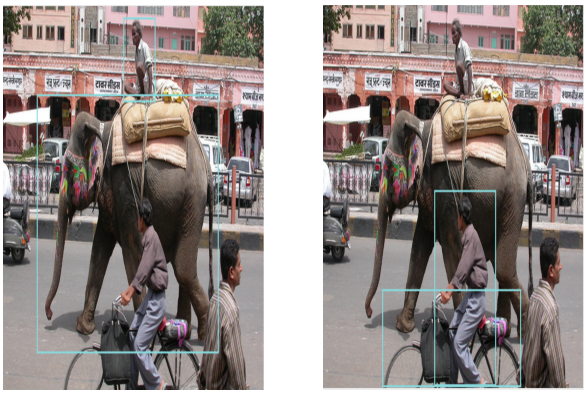
Model	Expl(S, R)	Impl(S, R)	Expl(O, R)	Impl(O, R)	Expl(R)	Impl(R)
BERT	61.9	14.4	59.9	27.0	**24.1**	**13.7**
FF+I	60.1	**68.6**	59.5	50.1	0	0
BERT+FF+I	**67.4**	59.8	**62.5**	**54.1**	24.0	**13.7**

Table 2: Experiments for unseen subjects, objects and relations (for both explicit and implicit relations). Spatial BERT gives better performance than BERT or FF+I for Impl(O, R) but not in Impl(S, R) potentially because the subject set is much sparser than the object set.



(a) man RIDING elephant (b) woman RIDING elephant

Figure 5: In this example from Visual Genome,suppose we have seen $(man, riding, elephant)$ in training. In the second image, we want to be able to predict *riding*, even if we have never seen the $(woman, riding)$ combination before



(a) man RIDING elephant (b) man RIDING bike

Figure 6: In this example from Visual Genome, suppose we have seen $(man, riding, elephant)$ in training. In the second image, we want to be able to predict *riding*, even if we have never seen the $(riding, bike)$ combination before

4.2. Analysis

Unseen Subject. For the explicit relations, we see that Spatial BERT performs much better than either model in isolation. This is potentially because the explicit relations are a smaller set and easier for BERT to predict and this in turn helps Spatial BERT. However, since the class of implicit relations are much larger and subtle BERT performs very poorly and spatial model by itself performs the best for the implicit relations.

Unseen Object. As shown in Table 2, for both the explicit and the implicit relations Spatial BERT gives the best performance although BERT by itself does not perform very well.

Unseen Relation. This task is possible because BERT is able to predict a much larger class of relations and using our GloVe-scoring metric we can decompose the scores of these relations across the known set of relations and this should bring the correct relations towards the top of the ranked list. We also relax the accuracy metric by counting a prediction as correct if the gold relation is in the top-5 predicted relations. In this setting, the spatial model by itself cannot predict any unseen relation and thus, BERT gives the best performance.

5. Discussion

We presented the task of spatial relation prediction and developed models that use position information, visual embeddings and word embeddings to predict relations. Further we show that combining this model with BERT helps for low resource settings and generalization over unseen subjects, objects and relations. However, the visual embeddings for entities which are used in our models are of limited use in some situations. For example, it is hard for

our models to distinguish between (*man*, *running*, *dog*) and (*man*, *walking*, *dog*) given (*man*, –, *dog*). In future we want to use more principled ways of incorporating image specific information to have an even more fine-grained spatial relation classification. We also want to develop an interactive system that does both the object position prediction task (Collell et al., 2018) and the spatial relation prediction task for a spatially-involved domain such as Blocks World (Bisk et al., 2016).

6. Acknowledgment

This work was supported by Contract W911NF-15-1-0461 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) and by Contract FA8750-19-2-0201 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

7. References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Baldridge, J., Bedrax-Weiss, T., Luong, D., Narayanan, S., Pang, B., Pereira, F., Soricut, R., Tseng, M., and Zhang, Y. (2018). Points, paths, and playscapes: Large-scale spatial language understanding tasks set in the real world. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 46–52.

Bisk, Y., Marcu, D., and Wong, W. (2016). Towards a dataset for human computer communication via grounded language acquisition. In *AAAI Workshop: Symbiotic Cognitive Systems*.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*.

Collell, G. and Moens, M.-F. (2018). Learning representations specialized in spatial knowledge: Leveraging language and vision. *Transactions of the Association of Computational Linguistics*, 6:133–144.

Collell, G., Van Gool, L., and Moens, M.-F. (2018). Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emami, A., De La Cruz, N., Trischler, A., Suleman, K., and Cheung, J. C. K. (2018). A knowledge hunting framework for common sense reasoning. *arXiv preprint arXiv:1810.01375*.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Kordjamshidi, P., Van Otterlo, M., and Moens, M.-F. (2010). Spatial role labeling: Task definition and annotation scheme. In *LREC*.

Kordjamshidi, P., Van Otterlo, M., and Moens, M.-F. (2011). Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):4.

Kordjamshidi, P., Rahgooy, T., and Manzoor, U. (2017). Spatial language understanding with multimodal graphs using declarative learning based programming. In *Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing*, pages 33–43.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Lin, X. and Parikh, D. (2016). Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pages 261–277. Springer.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sap, M., LeBras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y. (2018). Atomic: An atlas of machine commonsense for if-then reasoning. *arXiv preprint arXiv:1811.00146*.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Storks, S., Gao, Q., and Chai, J. Y. (2019). Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Wang, S. I., Ginn, S., Liang, P., and Manning, C. D. (2017). Naturalizing a programming language via interactive learning. *arXiv preprint arXiv:1704.06956*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

Xu, F. F., Lin, B. Y., and Zhu, K. Q. (2017). Commonsense locatednear relation extraction. *arXiv preprint arXiv:1711.04204*.