Learning Representations Specialized in Spatial Knowledge: Leveraging Language and Vision

Guillem Collell
Department of Computer Science
KU Leuven
3001 Heverlee, Belgium
gcollell@kuleuven.be

Marie-Francine Moens
Department of Computer Science
KU Leuven
3001 Heverlee, Belgium
sien.moens@cs.kuleuven.be

Abstract

Spatial understanding is crucial in many real-world problems, yet little progress has been made towards building representations that capture spatial knowledge. Here, we move one step forward in this direction and learn such representations by leveraging a task consisting in predicting continuous 2D spatial arrangements of objects given object-relationship-object instances (e.g., "cat under chair") and a simple neural network model that learns the task from annotated images. We show that the model succeeds in this task and, furthermore, that it is capable of predicting correct spatial arrangements for unseen objects if either CNN features or word embeddings of the objects are provided. The differences between visual and linguistic features are discussed. Next, to evaluate the spatial representations learned in the previous task, we introduce a task and a dataset consisting in a set of crowdsourced human ratings of *spatial similarity* for object pairs. We find that both CNN (convolutional neural network) features and word embeddings predict human judgments of similarity well and that these vectors can be further specialized in spatial knowledge if we update them when training the model that predicts spatial arrangements of objects. Overall, this paper paves the way towards building distributed *spatial representations*, contributing to the understanding of spatial expressions in language.

1 Introduction

Representing spatial knowledge is instrumental in any task involving text-to-scene conversion such as robot understanding of natural language commands (Guadarrama et al., 2013; Moratz and Tenbrink, 2006) or a number of robot navigation tasks. Despite recent advances in building specialized representations in domains such as sentiment analysis (Tang et al., 2014), semantic similarity/relatedness (Kiela et al., 2015) or dependency parsing (Bansal et al., 2014), little progress has been made towards building distributed representations (a.k.a. embeddings) specialized in spatial knowledge.

Intuitively, one may reasonably expect that the more attributes two objects share (e.g., *size*, *functionality*, etc.), the more likely they are to exhibit similar spatial arrangements with respect to other objects. Leveraging this intuition, we foresee that visual and linguistic representations can be spatially informative about unseen objects as they encode features/attributes of objects (Collell and Moens, 2016). For instance, without having ever seen an "elephant" before, but only a "horse", one would probably devise the "elephant" carrying the "human" than otherwise, just by considering their *size* attribute. Similarly, one can infer that a "tablet" and a "book" will show similar spatial patterns (usually on a table, in someone's hands, etc.) although they barely show any visual resemblance—yet they are similar in *size* and *functionality*. In this paper we systematically study how informative visual and linguistic features—in the form of convolutional neural network (CNN) features and word embeddings—are about the spatial behavior of objects.

An important goal of this work is to learn distributed representations specialized in spatial knowledge. As a vehicle to learn spatial representations,

Transactions of the Association for Computational Linguistics, vol. 6, pp. 133–144, 2018. Action Editor: Stefan Riezler.
Submission batch: 6/2017; Revision batch: 9/2017; Published 2/2018.

we leverage the task of predicting the 2D spatial arrangement for two objects under a relationship expressed by either a preposition (e.g., "below" or "on") or a verb (e.g., "riding", "jumping", etc.). For that, we make use of images where both objects are annotated with bounding boxes. For instance, in an image depicting (horse, jumping, fence) we reasonably expect to find the "horse" above the "fence". To learn the task, we employ a feed forward network that represents objects as continuous (spatial) features in an embedding layer and guides the learning with a distance-based supervision on the objects' coordinates. We show that the model fares well in this task and that by informing it with either word embeddings or CNN features it is able to output accurate predictions about unseen objects, e.g., predicting the spatial arrangement of (man, riding, bike) without having ever been exposed to a "bike" before. This result suggests that the semantic and visual knowledge carried by the visual and linguistic features correlates to a certain extent with the spatial properties of words, thus providing predictive power for unseen objects.

To evaluate the quality of the spatial representations learned in the previous task, we introduce a task consisting in a set of 1,016 human ratings of *spatial similarity* between object pairs. It is thus desirable for spatial representations that "spatially similar" objects (i.e., objects that are arranged *spatially similar* in most situations and relative to other objects) have similar embeddings. In these ratings we show, first, that both CNN features and word embeddings are good predictors of human judgments, and second, that these vectors can be further specialized in spatial knowledge if we update them by backpropagation when learning the model in the task of predicting spatial arrangements of objects.

The rest of the paper is organized as follows. In Sect. 2 we review related research. In Sect. 3 we describe two spatial tasks and a model. In Sect. 4 we describe our experimental setup. In Sect. 5 we present and discuss our results. Finally, in Sect. 6 we summarize our contributions.

2 Related Work

Contrary to earlier rule-based approaches to spatial understanding (Kruijff et al., 2007; Moratz and Tenbrink, 2006), Malinowski and Fritz (2014) propose a learning-based method that learns the parameters of "spatial templates" (or regions of acceptability of an object under a spatial relation) using a pooling approach. They show improved performance in image retrieval and image annotation (i.e., retrieving sentences given a query image) over previous rule-based systems and methods that rely on handcrafted templates. Contrary to us, they restrict to relationships expressed by explicit spatial prepositions (e.g., "on" or "below") while we also consider actions (e.g., "jumping"). Furthermore, they do not build spatial representations for objects.

Other approaches have shown the value of properly integrating spatial information into a variety of tasks. For example, Shiang et al. (2017) improve over the state-of-the-art object recognition by leveraging previous knowledge of object co-occurrences and relative positions of objects—which they mine from text and the web—in order to rank possible object detections. In a similar fashion, Lin and Parikh (2015) leverage common sense visual knowledge (e.g., object locations and co-occurrences) in two tasks: fill-in-the-blank and visual paraphrasing. They compute the likelihood of a scene to identify the most likely answer to multiple-choice textual scene descriptions. In contrast, we focus solely on spatial information rather than semantic plausibility. Moreover, our primary target is to build (spatial) representations. Alternatively, Elliott and Keller (2013) annotate geometric relationships between objects in images (e.g., they add an "on" link between "man" and "bike" in an image of a "man" "riding" a "bike") to better infer the action present in the image. For instance, if the "man" is next to the bike one can infer that the action "repairing" is more likely than "riding" in this image. Accounting for this extra spatial structure allows them to outperform bag-of-features methods in an image captioning task. In contrast with those who restrict to a small domain of 10 actions (e.g., "taking a photo", "riding", etc.), our goal is to generalize to any unseen/rare objects and actions by learning from frequent spatial configurations and objects, and critically, leveraging representations of objects. Recent work (Collell et al., 2018) tackles the research question of whether relative spatial arrangements can be predicted equally well from actions (e.g., "riding") than from spatial

prepositions (e.g., "below"), and how to interpret the learned weights of the network. In contrast, our research questions concern spatial representations. Crucially, none of the studies above have considered or attempted to learn distributed *spatial representations* of objects, nor studied how much spatial knowledge is contained in visual and linguistic representations.

The existence of quantitative, continuous *spatial representations* of objects has been formerly discussed, yet to our knowledge, not systematically investigated before. For instance, Forbus et al. (1991) conjectured that *"there is no purely qualitative, general purpose representation of spatial properties"*, further emphasizing that the quantitative component is strictly necessary.

It is also worth commenting on early work aimed at enhancing the understanding of natural spatial language such as the L_0 project (Feldman et al., 1996). In the context of this project, Regier (1996) proposed a connectionist model that learns to predict a few spatial prepositions ("above", "below", "left", "right", "in", "out", "on", and "off") from low resolution videos containing a limited set of toy objects (circle, square, etc.). In contrast, we consider an unlimited vocabulary of real-world objects, and we do not restrict to spatial prepositions but we include actions, as well. Hence, Regier's (1996) setting does not seem plausible to deal with actions given that, in contrast to the spatial prepositions that they use, which are mutually exclusive (an object cannot be "above" and simultaneously "below" another object), actions are not. In particular, actions exhibit large spatial overlap and, therefore, attempt to predict thousands of different actions from the relative locations of the objects seems infeasible. Additionally, Regier's (1996) architecture does not allow to meaningfully extract representations of *objects* from the visual input—which yields rather visual features.

Here, we propose an *ad hoc* setting for both, learning and evaluating spatial representations. In particular, instead of learning to predict spatial relations from visual input as in Regier's (1996) work, we learn the reverse direction, i.e., to map the relation (and two objects) to their visual spatial arrangement. By backpropagating the embeddings of the objects while learning the task, we enable learning spatial representations. As a core finding, we show in an *ad hoc* task, namely our collected human ratings of spatial similarity, that the learned features are more specialized in spatial knowledge than the CNN features and word embeddings that were used to initialize the parameters of the embeddings.

3 Tasks and Model

Here, we first describe the Prediction task and model that we use to learn the spatial representations. We subsequently present the spatial Similarity task which is employed to evaluate the quality of the learned representations.

3.1 Prediction Task

To evaluate the ability of a model or embeddings to learn spatial knowledge, we employ the task of predicting the spatial location of an Object ("O") relative to a Subject ("S") under a Relationship ("R"). Let $O^c = (O^c_x, O^c_y)$ denote the coordinates of the center ("c") of the Object's bounding box, where $O^c_x \in \mathbb{R}$ and $O^c_y \in \mathbb{R}$ are its x and y components. Let $O^b = (O^b_x, O^b_y)$ be one half of the vertical ($O^b_y \in \mathbb{R}$) and horizontal ($O^b_x \in \mathbb{R}$) sizes of the Object's bounding box ("b"). A similar notation applies to the Subject (i.e., S^c and S^b), and we denote model predictions with a hat $\widehat{O^c}$, $\widehat{O^b}$. The task is to learn a mapping from the structured textual **input** (Subject, Relation, Object)—abbreviated by (S, R, O)—to the **output** consisting of the Object's center coordinates O^c and its size O^b (see Fig. 1).

We notice that a "Subject" is not necessarily a syntactic subject but simply a convenient notation to accommodate the case where the Relationship (R) is an action (e.g., "riding" or "wearing"), while when R is a spatial preposition (e.g., "below" or "on") the Subject simply denotes the referent object. Similarly, the Object is not necessarily a direct object.[1]

3.2 Regression Model

Following the task above (Sect. 3.1), we consider a model (Fig. 1) that takes a triplet of words (S, R, O) as input and maps their one-

[1]We prefer to adhere to the terminology used to express entity-relationships in the Visual Genome dataset, but are aware of annotation schemes for spatial semantics (Pustejovsky et al., 2012). However, a one-to-one mapping of the Visual Genome terminology to these annotation schemes is not always possible.

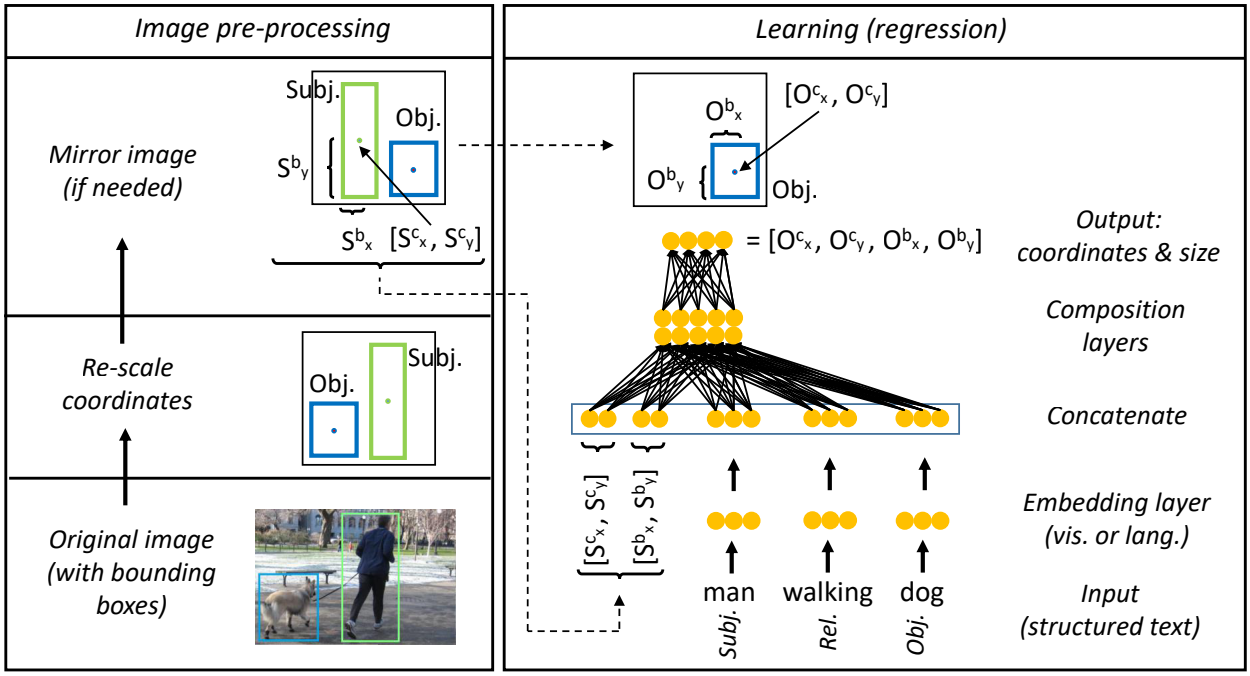

Figure 1: Overview of the model (right) and the image pre-processing setting (left).

hot[2] vectors w_S, w_R, w_O to d-dimensional dense vectors $w_S W_S, w_R W_R, w_O W_O$ via dot product with their respective *embedding* matrices $W_S \in \mathbb{R}^{d \times |V_S|}, W_R \in \mathbb{R}^{d \times |V_R|}, W_O \in \mathbb{R}^{d \times |V_O|}$, where $|V_S|, |V_R|, |V_O|$ are the vocabulary sizes. The embedding layer models our intuition that spatial properties of objects can be, to a certain extent, encoded with a vector of continuous features. In this work we test two types of embeddings, visual and linguistic. The next layer simply concatenates the three embeddings together with the Subject's size S^b and Subject center S^c. The inclusion of the Subject's size is aimed at providing a reference size to the model in order to predict the size of the Object O^b.[3] The resulting concatenated vector $[w_S W_S, w_R W_R, w_O W_O, S^c, S^b]$ is then fed into a hidden layer(s) which acts as a *composition function* for the triplet (S, R, O):

$$z = f(W_h[w_S W_S, w_R W_R, w_O W_O, S^c, S^b] + b_h)$$

where $f(\cdot)$ is the non-linearity and W_h and b_h the parameters of the layer. These "composition layers" allow to distinguish between e.g., (man, walks, horse) which is spatially distinct from (man, rides, horse). We find that adding more layers generally improves performance, so the output z above can simply be composed with more layers, i.e., $f(W_{h_2} z + b_{h_2})$. Finally, a linear output layer tries to match the ground truth targets $y = (O^c, O^b)$ using a mean squared error (MSE) loss function:

$$Loss(y, \hat{y}) = \|\hat{y} - y\|^2$$

where $\hat{y} = (\widehat{O^c}, \widehat{O^b})$ is the model prediction and $\|\cdot\|$ denotes the Euclidean norm. Critically, unlike CNNs, the model does not make use of the pixels (which are discarded during the image pre-processing (Fig. 1 and Sect. 3.2.1)), but learns exclusively from image coordinates, yielding a simpler model focused solely on spatial information.

[2] One-hot vectors (a.k.a. one-of-k), are sparse vectors with 0 everywhere except for a 1 at the position of the k-th word.

[3] We find that without inputing S^b to the model, it still learns to predict an "average size" for each Object due to the MSE penalty. However, to keep the design cleaner and intuitive, here we provide the size of the Subject to the model.

3.2.1 Image Pre-Processing

We perform the following pre-processing steps to the images before feeding them to the model.
(i) Normalize the image coordinates by the number of pixels of each axis (vertical and horizontal). This step guarantees that coordinates are independent of the resolution of the image and always lie within the $[0, 1] \times [0, 1]$ square, i.e., $S^c, O^c \in [0, 1]^2$.
(ii) Mirror the image (when necessary). We notice that the distinction between right and left is arbitrary in images since a mirrored image completely preserves its spatial meaning. For instance, a "man" "feeding" an "elephant" can be arbitrarily at either side of the "elephant", while a "man" "riding" an "elephant" cannot be either below or above the "elephant". This left/right arbitrariness has also been acknowledged in prior work (Singhal et al., 2003). Thus, to enable a more meaningful learning, we mirror the image when (and only when) the Object is at the left-hand side of the Subject.[4] The choice of leaving the Object always to the right-hand side is arbitrary and does not entail a loss of generality, i.e., we can consider left/right symmetrically reflected predictions as equiprobable. Mirroring provides thus a more realistic performance evaluation in the Prediction task and enables learning representations independent of the right/left distinction which is irrelevant for the spatial semantics.

3.3 Spatial Similarity Task

To evaluate how well our embeddings match human mental representations of spatial knowledge about objects, we collect ratings for 1,016 word pairs (w_1, w_2) asking annotators to rate them by their *spatial similarity*. That is, *objects that exhibit similar locations in most situations and are placed similarly relative to other objects would receive a high score, and lower otherwise.* For example (cap, sunglasses) would receive a high score as they are usually at the top of the human body, while following a similar logic, (cap, shoes) would receive a lower score. Our collected ratings establish the *spatial* counterpart to other existing similarity ratings such as *semantic* similarity (Silberer and Lapata, 2014), *visual* similarity (Silberer and Lapata, 2014) or general *relatedness* (Agirre et al., 2009). A few exemplars of ratings are shown in Tab. 1. Following standard practices (Pennington et al., 2014), we compute the prediction of similarity between two embeddings s_{w_1} and s_{w_2} (representing words w_1 and w_2) with their cosine similarity:

$$\cos(s_{w_1}, s_{w_2}) = \frac{s_{w_1} s_{w_2}}{\|s_{w_1}\| \|s_{w_2}\|}$$

We notice that this spatial Similarity task does not involve learning and its main purpose is to *evaluate* the quality of the representations learned in the Prediction task (Sect. 3.1) and the spatial informativeness of visual and linguistic features.

Word pair	Rating	Word pair	Rating
(snowboard, feet)	7.2	(horns, backpack)	1.8
(ears, eye)	8.3	(baby, bag)	7
(cockpit, table)	2.4	(hair, laptop)	1.8
(cap, hair)	9	(earring, racket)	2
(frisbee, food)	2.4	(ears, hat)	5.6

Table 1: Examples of our collected similarity ratings.

4 Experimental Setup

In this section we describe the experimental settings employed in the tasks and the model.

4.1 Visual Genome Data Set

We obtain our annotated data from Visual Genome (Krishna et al., 2017). This dataset contains 108,077 images and over 1.5M human-annotated object-relationship-object instances (S, R, O) with their corresponding boxes for the Object and Subject. We keep only those examples for which we have embeddings available (see Sect. 4.3). This yields $\sim$1.1M instances of the form (S, R, O), 7,812 unique image objects and 2,214 unique Relationships (R) for our *linguistic* embeddings; and $\sim$920K (S, R, O) instances, 4,496 unique image objects and 1,831 unique Relationships for our *visual* embeddings. We notice that visual representations do not exist for Relationships R (i.e., either prepositions or verbs) and therefore we only require visual embeddings for the pair (S, O) instead of the complete triplet (S, R, O) required in language. Notice that since we

[4]The only conflicting case for the "mirroring" transformation is when the Relationship (R) is either "left" or "right," e.g., (man, left of, car), yet these only account for an insignificant proportion of instances ($< 0.1\%$) and thus we leave them out.

do not restrict to any particular domain (e.g., furniture or landscapes) the combinations (S, R, O) are markedly sparse, which makes learning our Prediction task especially challenging.

4.2 Evaluation Sets in the Prediction Task

In the Prediction task, we consider the following subsets of Visual Genome (Sect. 4.1) for evaluation purposes:

(i) Original set: a test split from the original data which contains *instances* unseen at training time. That is, the test *combinations* (S, R, O) might have been seen at training time, yet in different *instances* (e.g., in different images). This set contains a large number of noisy combinations such as (people, walk, funny) or (metal, white, chandelier).

(ii) Unseen Words set: We randomly select a list of 25 objects (e.g., "wheel", "camera", "elephant", etc.) among the 100 most frequent objects in Visual Genome.[5] We choose them among the most frequent ones in order to avoid meaningless objects such as "gate 2", "number 40" or "2:10 pm" which are not infrequent in Visual Genome. We then take all instances of combinations that contain any of these words, yielding $\sim$ 123K instances. For example, since "cap" is in our list, (girl, wears, cap) is included in this set. When we enforce "*unseen*" conditions, we remove all these instances from the training set, using them only for testing.

4.3 Visual and Linguistic Features

As our *linguistic representations*, we employ 300-dimensional GloVe vectors (Pennington et al., 2014) trained on the Common Crawl corpus with 840B-tokens and a 2.2M words vocabulary.[6]

We use the publicly available *visual representations* from Collell et al. (2017).[7] They extract 128-dimensional visual features with the forward pass of a VGG-128 (Visual Geometry Group) CNN model (Chatfield et al., 2014) pre-trained in ImageNet (Russakovsky et al., 2015). The representation of a word is the averaged feature vector (centroid) of all images in ImageNet for this concept. They only keep words with at least 50 images available. We notice that although we employ visual features from an external source (ImageNet), these could be alternatively obtained in the Visual Genome data—although ImageNet generally provides a larger number of images per concept.

4.4 Method Comparison

We consider two types of models, those that update the parameters of the embeddings (***U*** $\sim$ "Update") and those that keep them fixed (***NU*** $\sim$ "No Update") when learning the Prediction task. For each type (*U* and *NU*) we consider two conditions, embeddings initialized with pre-trained vectors (***INI***) and random embeddings (***RND***) randomly drawn from a component-wise normal distribution of mean and standard deviation equal to those of the original embeddings. For example, *U-RND* corresponds to a model with updated, random embeddings. For the *INI* methods we also add a subindex indicating whether the embeddings are visual (***vis***) or linguistic (***lang***), as described in Sect. 4.3.[8] For the *NU* type we additionally consider *one-hot* embeddings (***1H***). We also include a control method (***rand-pred***) that outputs random uniform predictions.

4.5 Implementation Details and Validation

To validate results in our Prediction task we employ a 10-fold cross-validation (CV) scheme. That is, we split the data into 10 parts and employ 90% of the data for training and 10% for testing. This yields 10 embeddings (for each "*U*" method), which are then evaluated in our Similarity task. In both tasks, we report results averaged across the 10 folds.

Model hyperparameters are first selected by cross-validation in 10 initial splits and results are reported in 10 new splits. All models employ a learning rate of 0.0001 and are trained for 10 epochs by backpropagation with the RMSprop optimizer. The dimensionality of the embeddings is the original one, i.e., d=300 for GloVe and d=128 for VGG-128 (Sect. 4.3), which is preserved for the random-

[5]The complete list of objects is: [leaves, foot, wheel, t-shirt, ball, handle, skirt, stripe, trunk, face, camera, socks, tail, pants, elephant, ear, helmet, vest, shoe, eye, coat, skateboard, apple, cap, motorcycle].

[6]http://nlp.stanford.edu/projects/glove

[7]http://liir.cs.kuleuven.be/software.php

[8]Given that visual representations are not available for the Relationships (i.e., verbs and prepositions), the models with *vis* embeddings employ one-hot embeddings for the Relationships and visual embeddings for Object and Subject. This is a rather neutral choice that enables the *vis* models to use Relationships.

embedding methods *RND* (Sect. 4.4). Models employ 2 hidden layers with 100 Rectified Linear Units (ReLu), followed by an output layer with a linear activation. Early stopping is employed as a regularizer. We implement our models with Keras deep learning framework in Python 2.7 (Chollet and others, 2015).

4.6 Spatial Similarity Task

To build the word pairs, we randomly select a list of objects from Visual Genome and from these we randomly chose 1,016 non-repeated word pairs (w_1, w_2). Ratings are collected with the Crowdflower[9] platform and correspond to averages of at least 5 reliable annotators[10] that provided ratings in a discrete scale from 1 to 10. The median similarity rating is 3.3 and the mean variance between annotators per word pair is $\sim$1.2.

4.7 Evaluation Metrics

4.7.1 Prediction Task

We evaluate model predictions with the following metrics.

(I) Regression metrics.
(i) Mean Squared Error (MSE) between the predicted $\hat{y} = (\widehat{O^c}, \widehat{O^b})$ and the true $y = (O^c, O^b)$ Object center coordinates and Object size. Notice that since O^c, O^b are within $[0, 1]^2$, the MSE is easily interpretable, ranging between 0 and 1.
(ii) Pearson Correlation (r) between the predicted $\widehat{O^c}$ and the true O^c Object center coordinates. We consider the vertical ($\mathbf{r}_y$) and horizontal ($\mathbf{r}_x$) components separately (i.e., O^c_x and O^c_y).
(iii) Coefficient of Determination ($\mathbf{R^2}$) of the predictions $\hat{y} = (\widehat{O^c}, \widehat{O^b})$ and the target $y = (O^c, O^b)$. R^2 is employed to evaluate goodness of fit of a regression model and is related to the percentage of variance of the target explained by the predictions. The best possible score is 1 and it can be arbitrarily negative for bad predictions. A model that outputs either random or constant predictions would obtain scores close to 0 and exactly 0 respectively.

[9]https://www.crowdflower.com/

[10]Reliable annotators are those with performance over 70% in the test questions (16 in our case) that the crowdsourcing platform allows us to introduce in order to test annotators' accuracy.

(II) Classification. Additionally, given the semantic distinction between the vertical and horizontal axis noted above (Sect. 3.2.1), we consider the classification problem of predicting above/below relative locations. That is, if the predicted y-coordinate for the Object center $\widehat{O^c_y}$ falls *below* the y-coordinate of the Subject center S^c_y and the actual Object center O^c_y is below the Subject center S^c_y, we count it as a correct prediction, and as incorrect otherwise. Likewise for *above* predictions. We compute both macro-averaged[11] accuracy ($\mathbf{acc}_y$) and macro-averaged F1 ($\mathbf{F1}_y$) metrics.

(III) Intersection over Union (IoU). We consider the bounding box overlap (IoU) from the VOC detection task (Everingham et al., 2015): IoU $= area(\widehat{B_O} \cap B_O)/area(\widehat{B_O} \cup B_O)$ where $\widehat{B_O}$ and B_O are predicted and ground truth Object boxes respectively. A prediction is counted as correct if the IoU is larger than 50%. Crucially, we notice that our setting and results are not comparable to object detection as we employ text instead of images as input and thus we cannot leverage the pixels to locate the Object, unlike in detection.

4.7.2 Similarity Task

Following standard practices (Pennington et al., 2014), the performance of the predictions of (cosine) similarity from the embeddings (described in Sect. 3.3) is evaluated with the Spearman correlation ρ against the crowdsourced human ratings.

5 Results and Discussion

We consider the notation of the methods from Sect. 4.4 and the evaluation subsets described in Sect. 4.2 for the Prediction task. To test statistical significance we employ a Friedman rank test and post hoc Nemeny tests on the results of the 10 folds.

5.1 Prediction Task

Table 2 shows that the *INI* and *RND*[12] methods perform similarly in the *Original* test set, arguably be-

[11]Macro-averaged accuracy equals to the average of per-class accuracies, with classes {above, below}. Similarly for F1.

[12]For simplicity, we do not add any subindex (vis or $lang$) to *RND*, yet these vectors are drawn from two different distributions, i.e., from either vis or $lang$ embeddings (Sect. 4.4). Additionally, results tables show two blocks of methods since vis and $lang$ do not share all the instances (see Sect. 4.1).

cause a large part of the learning takes place in the parameters of the layers subsequent to the embedding layer. However, in the next section we show that this is no longer the case when *unseen* words are present. We also observe that the one-hot embeddings *NU-1H* perform slightly better than the rest of methods when no *unseen* words are present (Tab. 2 and Tab. 3 right).

	MSE	R^2	acc_y	$F1_y$	r_x	r_y	IoU
U-INI$_{lang}$	0.011	0.654	0.773	0.773	0.849	0.832	0.283
U-RND$_{lang}$	0.011	0.646	0.770	0.770	0.847	0.827	0.279
NU-INI$_{lang}$	0.011	0.651	0.770	0.770	0.848	0.829	0.275
NU-RND$_{lang}$	0.011	0.636	0.766	0.766	0.845	0.822	0.268
NU-1H	**0.010**	**0.659**	**0.777**	**0.778**	**0.850**	**0.833**	**0.297**
rand-pred	0.794	-27.61	0.533	0.516	0.000	0.001	0.010
U-INI$_{vis}$	**0.011**	0.627	0.766	0.766	0.841	0.820	0.266
U-RND$_{vis}$	0.012	0.612	0.762	0.762	0.836	0.810	0.244
NU-INI$_{vis}$	0.012	0.611	0.765	0.763	0.837	0.813	0.246
NU-RND$_{vis}$	0.012	0.607	0.767	0.766	0.835	0.808	0.237
NU-1H	**0.011**	**0.657**	**0.788**	**0.788**	**0.848**	**0.833**	**0.308**
rand-pred	0.789	-27.51	0.534	0.519	0.000	0.000	0.010

Table 2: Results in the **Original** test set (Sect. 4.2). Boldface indicates best performance within the corresponding block of methods ($lang$ above, and vis below).

It is also worth noting that the results of the Prediction task are, in fact, conservative. First, the *Original* test data contains a considerable number of meaningless (e.g., (giraffe, a, animal)), and irrelevant combinations (e.g., (clock, has, numbers) or (sticker, identifies, apple)). Second, even when only meaningful examples are considered, we are inevitably penalizing for plausible predictions. For instance, in (man, watching, man) we expect both men to be reasonably separated on the x-axis yet the one with the highest y coordinate is generally not predictable as it depends on their height and their distance to the camera. This yields above/below classification performance and correlations. Regardless, all methods (except *rand-pred*) exhibit reasonably high performance in all measures.

5.1.1 Evaluation on Unseen Words

Table 3 evidences that both visual and linguistic embeddings (*INI*$_{vis}$ and *INI*$_{lang}$) significantly outperform their random-embedding counterparts *RND* by a large margin when *unseen* words are present. The improvement occurs for both, updated (*U*) and non-updated (*NU*) embeddings—although it is expected that the updated methods perform slightly worse than the non-updated ones since the original embeddings will have "moved" during training and therefore an *unseen* embedding (which has not been updated) might no longer be close to other semantically similar vectors in the updated space.

Besides statistical significance, it is worth mentioning that the *INI* methods consistently outperformed both their *RND* counterparts and *NU-1H* in each of the 10 folds (not shown here) by a steadily large margin. In fact, results are markedly stable across folds, in part due to the large size of the training and test sets ($>$ 0.9M and $>$ 120K examples respectively). Additionally, to ensure that "*unseen*" results are not dependent on our particular list of objects, we repeated the experiment with two additional lists of randomly selected objects, obtaining very similar results.

Remarkably, the *INI* methods experience only a small performance drop under *unseen* conditions (Tab. 3, left) compared to when we allow them to train with these words (Tab. 3, right), and this difference might be partially attributed to the reduction of the training data under "*unseen*" conditions, where at least 10% of the training data are left out.

Altogether, these results on unseen words show that semantic and visual similarities between concepts, as encoded by word and visual embeddings, can be leveraged by the model in order to predict spatial knowledge about unseen words.[13]

5.1.2 Qualitative Insight

Visual inspection of model predictions is instructive in order to gain insight on the spatial informativeness of visual and linguistic representations on *unseen* words. Figure 2 shows heat maps of low (black) and high (white) probability regions for the objects. The "heat" for the Object is assumed to be normally distributed with mean (μ) equal to the predicted Object center $\widehat{O^c}$ and standard deviation (σ) equal to the predicted Object size $\widehat{O^b}$ (assuming independence of the x and y components, which yields the product of two Gaussians, one for each component x and y). The "heat" for the Subject is computed similarly, although with μ and σ equal to the

[13] In this Prediction task we have additionally considered the concatenation of visual and linguistic representations (not shown), which did not show any relevant improvement over the unimodal representations.

	Unseen words condition							Seen words condition						
	MSE	R^2	acc_y	$F1_y$	r_x	r_y	IoU	MSE	R^2	acc_y	$F1_y$	r_x	r_y	IoU
U-INI$_{lang}$	0.011$^{*\diamond}$	0.584$^{*\diamond}$	0.712$^{*\diamond}$	0.710$^{*\diamond}$	0.877$^{*\diamond}$	0.770*	0.131$^{*\diamond}$	**0.007**	0.736*	0.810	0.810	**0.901**	0.876	0.223
U-RND$_{lang}$	0.015	0.422	0.603	0.601	0.863	0.624	0.090	**0.007**	0.730	0.806	0.806	0.899	0.874	0.223
NU-INI$_{lang}$	**0.009**$^{*\diamond}$	**0.663**$^{*\diamond}$	**0.770**$^{*\diamond}$	**0.770**$^{*\diamond}$	**0.888**$^{*\diamond}$	**0.835**$^{*\diamond}$	**0.164**$^{*\diamond}$	**0.007**	0.734*	0.805	0.805	0.900	0.875	0.221
NU-RND$_{lang}$	0.016	0.405	0.600	0.598	0.864	0.617	0.101	**0.007**	0.721	0.803	0.803	0.898	0.871	0.212
NU-1H	0.015	0.465	0.608	0.607	0.867	0.642	0.098	**0.007**	**0.740**	**0.814**	**0.813**	**0.901**	**0.877**	**0.243**
rand-pred	0.843	-38.32	0.524	0.501	0.000	0.000	0.012	0.845	-38.42	0.524	0.500	-0.002	-0.001	0.012
U-INI$_{vis}$	**0.010**$^{*\diamond}$	0.599$^{*\diamond}$	0.775$^{*\diamond}$	0.774$^{*\diamond}$	**0.887**$^{*\diamond}$	0.801$^{*\diamond}$	**0.123**$^{*\diamond}$	0.007	0.726	0.816	0.816	0.904	0.874	0.200
U-RND$_{vis}$	0.017	0.360	0.581	0.578	0.867	0.513	0.082	0.008	0.711	0.812	0.811	0.901	0.864	0.174
NU-INI$_{vis}$	**0.010**$^{*\diamond}$	**0.602**$^{*\diamond}$	**0.777**$^{*\diamond}$	**0.775**$^{*\diamond}$	**0.887**$^{*\diamond}$	**0.803**$^{*\diamond}$	**0.123**$^{*\diamond}$	0.007	0.711	0.817	0.815	0.902	0.868	0.186
NU-RND$_{vis}$	0.017	0.366	0.574	0.572	0.867	0.536	0.085	0.008	0.706	0.820	0.819	0.901	0.862	0.171
NU-1H	0.015	0.437	0.618	0.617	0.867	0.601	0.078	**0.006**	**0.760**	**0.841**	**0.841**	**0.910**	**0.885**	**0.256**
rand-pred	0.840	-40.34	0.524	0.507	-0.001	0.000	0.012	0.840	-40.37	0.524	0.507	-0.002	-0.001	0.012

Table 3: Results in the **Unseen Words** set (Sect. 4.2). Left table: results of enforcing "*unseen*" conditions, i.e., leaving out all words of the *Unseen Words* set from our training data. Right table: the models are evaluated in the same set but we allow them to train with the words from this set. Asterisks (*) in an *INI* method indicate significantly better performance ($p < 0.05$) than its RND counterpart (i.e., *U-INI*$_{emb\_type}$ is compared against *U-RND*, and *NU-INI*$_{emb\_type}$ against *NU-RND*). Diamonds ($^\diamond$) indicate significantly better performance than *NU-1H*.

actual Subject center S^c and size S^b, respectively.

The *INI* methods in Figure 2 illustrate the contribution of the embeddings to the spatial understanding of *unseen* objects. In general, both visual and linguistic embeddings enabled predicting meaningful spatial arrangements, yet for the sake of space we have only included three examples where: vis performs better than $lang$ (third column), where $lang$ performs better than vis (second column), and where both perform well (first column). We notice that the embeddings enable the model to infer that e.g., since "camera" (*unseen*) is similar to "camcorder" (seen at training time), both must behave spatially similarly. Likewise, the embeddings enable predicting correctly the relative sizes of unseen objects. We also observe that when the embeddings are not informative enough, model predictions become less accurate. For instance, in *NU-INI*$_{lang}$, some unrelated objects (e.g., "ipod") have embeddings similar to "apple", and analogously for *NU-INI*$_{vis}$ and "tail". We finally notice that predictions on *unseen* objects using random embeddings (*RND*) are markedly bad.

5.2 Spatial Similarity Task

Table 4 shows the results of evaluating the embeddings, including those learned in the Prediction task, against the human ratings of *spatial similarity* (Sect. 3.3). Hence, only the "updated" methods (*U*) are shown and we additionally include the concatenation of visual and linguistic embeddings *CONC*$_{GloVe+VGG\text{-}128}$ and the concatenation of the corresponding updated embeddings *CONC*$_{U\text{-}INI_{lang}+U\text{-}INI_{vis}}$.

	LANG	V&L
GloVe	0.543	0.535
VGG-128	-	0.459
CONC$_{GloVe+VGG\text{-}128}$	-	0.582
U-INI$_{lang}$	$0.557 \pm 0.0015^*$	$0.558 \pm 0.002^*$
U-INI$_{vis}$	-	$0.48 \pm 0.0012^*$
CONC$_{U\text{-}INI_{lang}+U\text{-}INI_{vis}}$	-	$0.6 \pm 0.0015^*$
U-RND	0.15 ± 0.0075	0.174 ± 0.0078
# word pairs	1016	839

Table 4: Spearman correlations between model predictions and human ratings. Standard errors across folds are shown for the methods that involve learning (second block). Columns correspond to the word pairs for which both embeddings (*vis* and *lang*) are available (V&L) and those for which only the linguistic embeddings are available (LANG). Asterisk (*) indicates significant improvement ($p < 0.05$) of a *U-INI* method of the second block (*U-INI*$_{vis}$ and *U-INI*$_{lang}$) over its corresponding untrained embedding (i.e., *VGG-128* or *GloVe* respectively) from the first block.

The first thing to notice in Tab. 4 is that both visual and linguistic embeddings show good correlations with human spatial ratings ($\rho > 0.45$ and $\rho > 0.53$ respectively), suggesting that visual and linguistic features carry significant knowledge about

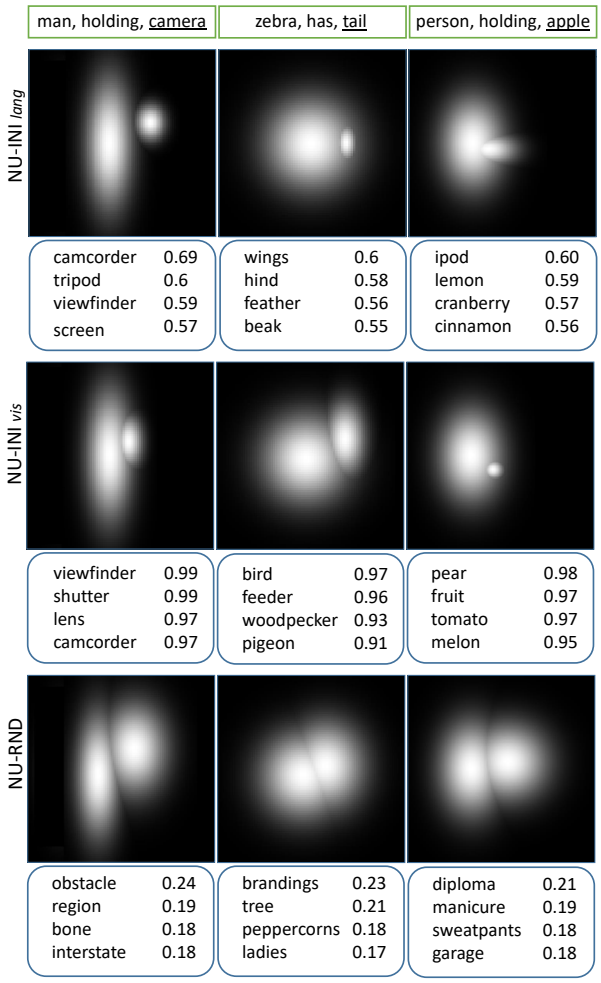

Figure 2: Heat maps of predictions of the *NU-INI*$_{lang}$, *U-INI*$_{vis}$ and *NU-RND* methods. The *unseen* Objects are underlined (top of the image) and their corresponding four (cosine-based) nearest neighbors are shown below with their respective cosine similarities.

spatial properties of objects. In particular, linguistic features seem to be more spatially informative than visual features.

Crucially, we observe a significant improvement of the *U-INI*$_{vis}$ over the original visual vectors (*VGG-128*) ($p < 0.05$) and of the *U-INI*$_{lang}$ over the original linguistic embeddings (*GloVe*) ($p < 0.05$), which evidence the effectiveness of training in the Prediction task as a method to further specialize embeddings in spatial knowledge. It is worth mentioning that these improvements are consistent in each of the 10 folds (not shown here) and markedly stable (see standard errors in Tab. 4).

We additionally observe that the concatenation of visual and linguistic embeddings *CONC*$_{GloVe+VGG-128}$ outperforms all unimodal embeddings by a margin, suggesting that the fusion of visual and linguistic features provides a more complete description of spatial properties of objects. Remarkably, the improvement is even larger for the concatenation of the embeddings updated during training *CONC*$_{U\text{-}INI_{lang}+U\text{-}INI_{vis}}$, which obtains the highest performance overall.

Figure 3 illustrates the progressive specialization of our embeddings in spatial knowledge as we train them in our Prediction task. We notice that all embeddings improve, yet *U-INI*$_{lang}$ seem to worsen their quality when we over-train them—likely due to overfitting, as we do not use any regularizer besides early stopping. We also observe that although the random embeddings (*RND*) are the ones that benefit the most from the training, their performance is still far from that of *U-INI*$_{vis}$ and *U-INI*$_{lang}$, suggesting the importance of visual and linguistic features to represent spatial properties of objects.

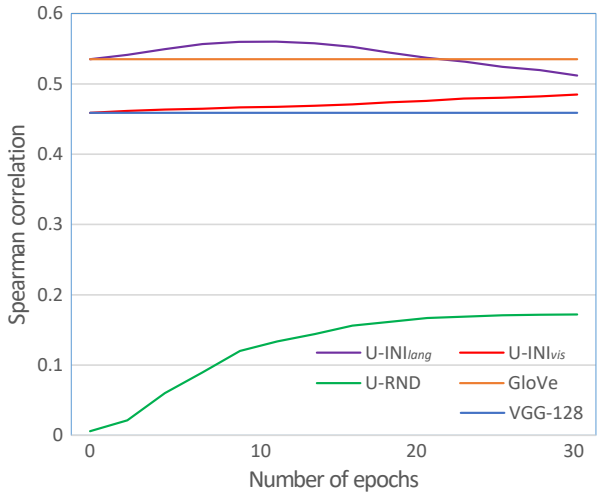

Figure 3: Correlation between human ratings and embedding cosine similarities at each number of epochs.

It is relevant to mention that in a pilot study we crowdsourced a different list of 1,016 object pairs where we employed 3 instead of 5 annotators per row. Results stayed remarkably consistent with those presented here—the improvement for the updated embeddings was in fact even larger.

Limitations of the current approach and future work In order to keep the design clean in this first paper on distributed *spatial representations* we employ a fully supervised setup. However, we notice that methods to automatically parse images (e.g., object detectors) and sentences are available.

A second limitation is the 2D simplification of the actual 3D world that our approach and the current spatial literature generally employs. Even though methods that infer 3D structure from 2D images exist, this is beyond the scope of this paper which shows that a 2D treatment already enhances the learned spatial representations. It is also worth noting that the proposed regression setting trivially generalizes to 3D if suitable data are available, and in fact, we believe that the learned representations could further benefit from such extension.

6 Conclusions

Altogether, this paper sheds light on the problem of learning distributed *spatial representations* of objects. To learn spatial representations we have leveraged the task of predicting the continuous 2D relative spatial arrangement of two objects under a relationship, and a simple embedding-based neural model that learns this task from annotated images. In the same Prediction task we have shown that both word embeddings and CNN features endow the model with great predictive power when is presented with unseen objects. Next, in order to assess the spatial content of distributed representations, we have collected a set of 1,016 object pairs rated by *spatial similarity*. We have shown that both word embeddings and CNN features are good predictors of human spatial judgments. More specifically, we find that word embeddings ($\rho = 0.535$) tend to perform better than visual features ($\rho \sim 0.46$), and that their combination ($\rho \sim 0.6$) outperforms both modalities separately. Crucially, in the same ratings we have shown that by training the embeddings in the Prediction task we can further specialize them in spatial knowledge, making them more akin to human spatial judgments. To benchmark the task, we make the Similarity dataset and our trained spatial representations publicly available.[14]

Lastly, this paper contributes to the automatic understanding of spatial expressions in language. The lack of common sense knowledge has been recurrently argued as one of the main reasons why machines fail at exhibiting more "human-like" behavior in tasks (Lin and Parikh, 2015). Here, we have provided a means of compressing and encoding such common sense spatial knowledge about objects into distributed representations, further showing that these specialized representations correlate well with human judgments. In future work, we will also explore the application of our trained spatial embeddings in extrinsic tasks in which representing spatial knowledge is essential such as robot navigation or robot understanding of natural language commands (Guadarrama et al., 2013; Moratz and Tenbrink, 2006). Robot navigation tasks such as assisting people with special needs (blind, elderly, etc.) are in fact becoming increasingly necessary (Ye et al., 2015) and require great understanding of spatial language and spatial connotations of objects.

Acknowledgments

This work has been supported by the CHIST-ERA EU project MUSTER[15] and by the KU Leuven grant RUN/15/005. We additionally thank our anonymous reviewers for their insightful comments which helped to improve the overall quality of the paper, and the action editors for their helpful assistance.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-Based Approaches. In *NAACL*, pages 19–27. ACL.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring Continuous Word Representations for Dependency Parsing. In *ACL*, pages 809–815.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *BMVC*.

François Chollet et al. 2015. Keras. `https://github.com/fchollet/keras`.

Guillem Collell and Marie-Francine Moens. 2016. Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences between Visual and

[14] https://github.com/gcollell/spatial-representations

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

Linguistic Representations. In *COLING*, pages 2807–2817. ACL.

Guillem Collell, Teddy Zhang, and Marie-Francine Moens. 2017. Imagined Visual Representations as Multimodal Embeddings. In *AAAI*, pages 4378–4384. AAAI.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates. In *AAAI*. AAAI.

Desmond Elliott and Frank Keller. 2013. Image Description Using Visual Dependency Representations. In *EMNLP*, volume 13, pages 1292–1302.

Mark Everingham, S.M. Ali Eslami, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. 2015. The PASCAL Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Jerome Feldman, George Lakoff, David Bailey, Srini Narayanan, Terry Regier, and Andreas Stolcke. 1996. L_0-The First Five Years of an Automated Language Acquisition Project. *Integration of Natural Language and Vision Processing*, 10:205.

Kenneth D. Forbus, Paul Nielsen, and Boi Faltings. 1991. Qualitative spatial reasoning: The CLOCK project. *Artificial Intelligence*, 51(1-3):417–471.

Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, Trevor Darrell, et al. 2013. Grounding Spatial Relations for Human-Robot Interaction. In *IROS*, pages 1640–1647. IEEE.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing Word Embeddings for Similarity or Relatedness. In *EMNLP*, pages 2044–2048.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated Dialogue and Spatial Organization: What, Where and Why? *International Journal of Advanced Robotic Systems*, 4(1):16.

Xiao Lin and Devi Parikh. 2015. Don't Just Listen, Use your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks. In *CVPR*, pages 2984–2993.

Mateusz Malinowski and Mario Fritz. 2014. A Pooling Approach to Modelling Spatial Relations for Image Retrieval and Annotation. *arXiv preprint arXiv:1411.5190v2*.

Reinhard Moratz and Thora Tenbrink. 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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543.

James Pustejovsky, Jessica Moszkowicz, and Marc Verhagen. 2012. A Linguistically Grounded Annotation Language for Spatial Information. *Traitement Automatique des Langues*, 53(2):87–113.

Terry Regier. 1996. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. MIT Press.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sz-Rung Shiang, Stephanie Rosenthal, Anatole Gershman, Jaime Carbonell, and Jean Oh. 2017. Vision-Language Fusion for Object Recognition. In *AAAI*, pages 4603–4610. AAAI.

Carina Silberer and Mirella Lapata. 2014. Learning Grounded Meaning Representations with Autoencoders. In *ACL*, pages 721–732.

Amit Singhal, Jiebo Luo, and Weiyu Zhu. 2003. Probabilistic Spatial Context Models for Scene Content Understanding. In *CVPR*, volume 1, pages 235–241. IEEE.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *ACL*, pages 1555–1565.

Cang Ye, Soonhac Hong, and Amirhossein Tamjidi. 2015. 6-DOF Pose Estimation of a Robotic Navigation Aid by Tracking Visual and Geometric Features. *IEEE Transactions on Automation Science and Engineering*, 12(4):1169–1180.