What Goes Into A Word: Generating Image Descriptions With Top-Down Spatial Knowledge

Mehdi Ghanimifard **Simon Dobnik**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science (FLoV)
University of Gothenburg, Sweden
{mehdi.ghanimifard,simon.dobnik}@gu.se

Abstract

Generating grounded image descriptions requires associating linguistic units with their corresponding visual clues. A common method is to train a decoder language model with attention mechanism over convolutional visual features. Attention weights align the stratified visual features arranged by their location with tokens, most commonly words, in the target description. However, words such as spatial relations (e.g. *next to* and *under*) are not directly referring to geometric arrangements of pixels but to complex geometric and conceptual representations. The aim of this paper is to evaluate what representations facilitate generating image descriptions with spatial relations and lead to better grounded language generation. In particular, we investigate the contribution of four different representational modalities in generating relational referring expressions: (i) (pre-trained) convolutional visual features, (ii) spatial attention over visual features, (iii) top-down geometric relational knowledge between objects, and (iv) world knowledge captured by contextual embeddings in language models.

1 Introduction

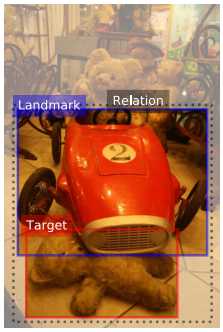
Spatial recognition and reasoning are essential bases for visual understanding. Automatically generating descriptions of scenes involves both recognising objects and their spatial configuration. This project follows up on recent attempts to improve language generation and understanding in terms of using spatial modules in the fusion of vision and language (Xu et al., 2015; Johnson et al., 2016; Lu et al., 2017; Hu et al., 2017; Anderson et al., 2018) (see also Section 6).

Generating spatial descriptions is an important part of the image description task which requires several types of knowledge obtained from different modalities: (i) invariant visual clues for object identification, (ii) geometric configuration of the scene representing relations between objects relative to the size of the environment (iii) object-specific functional relations that capture interaction between them and are formed by our knowledge of the world for example *an umbrella is over a man* is true if the referring umbrella serves its function, protecting the man from the rain (Coventry et al., 2001), and (iv) for projective relations (e.g. "to the left of" and "above") but not topological relations (e.g. "close" and "at"), the frame of reference which can be influenced from other modalities such as scene attention and dialogue interaction (Dobnik et al., 2015). Work in cognitive psychology (Logan, 1994, 1995) argues that while object identification may be pre-attentive, identification of spatial relations is not and is accomplished by a top-down mechanisms of attention after the objects have been identified. It is also the case that we do not identify all possible relations between objects but only those that are attended by such top-down mechanisms considering different kinds of high-level knowledge.

Experiments on training neural recurrent language models in a bottom-up fashion from data[1] demonstrated that spatial relations are frequently not learned to be grounded in visual inputs (Lu et al., 2017; Tanti et al., 2018a; Ghanimifard and Dobnik, 2018) which has been attributed to the design choices of these models that primarily focus on identification of objects (Kelleher and Dobnik, 2017). Therefore, targeted integration of different modalities is required to capture the properties from (i) to (iv). We can do this top-down (Anderson et al., 2018; Hu et al., 2017; Liu et al., 2017). However, it is not immediately obvious *what* kind of top-down spatial knowledge will benefit the bottom-up models most. Therefore, in this paper we investigate the integration of different kind of

[1] A bottom-up learning acquires higher level representations from examples of local features rather than using an external procedure to extract them. See also Section 6.

Proceedings of The 12th International Conference on Natural Language Generation, pages 540–551,
Tokyo, Japan, 28 Oct - 1 Nov, 2019.



⟨ "teddy bear", "partially under", "go cart"⟩

Figure 1: ⟨TARGET, RELATION, LANDMARK⟩ annotation of bounding boxes in VisualGenome 2318741[a]

[a]RaSeLaSeD_Il_Pinguino (2008): CC BY-SA 2.0.

top-down spatial knowledge beyond object localisation represented as features with the bottom-up neural language model.

The paper is organised as follows. In Section 2, we discuss how spatial descriptions are constructed and what components are required to generate descriptions. In Section 3, the neural networks' design is explained. In Section 4, we explain what dataset is used for this study, what preprocessing was applied on it and how the models are trained. Then the experiments and evaluation results are presented in Section 5. The related work in relation to our methods and findings is discussed in Section 6. The conclusion is given in Section 7.

2 Generating Spatial Descriptions

When describing a scene, there are several ways to construct spatial descriptions referring to objects and places and their relation with each other. A spatial description has three parts: a TARGET and a LANDMARK referring to objects or places and a RELATION denoting the location of the target in relation to the landmark (Logan and Sadler, 1996).[2] These are in the example in Figure 1 as follows:

There is *a teddy bear* *partially under* *a go cart*.
TARGET RELATION LANDMARK

Therefore generating such description requires (a) identification of objects and their locations: the target is what we want to describe and the landmark is what we will relate the target to; the salience of the landmark is important for the hearer. (b) Grounding of the relation in geometric space: the spatial relation is expressed relative to the landmark which grounds a 3-dimensional coordinate system; furthermore, for projective relations, the coordinate system is aligned with the orientation of the external viewpoint which determines the frame of reference (Maillat, 2003). (Viewpoint may also be the landmark object itself in which case the coordinate system is oriented in the same way as the landmark). (c) Grounding in function: a spatial relation may be selected also based on the functional properties between target and landmark objects, e.g. the difference between "*the teapot is over the cup*" and "*the teapot is above the cup*" (Coventry et al., 2001).

Generating spatial descriptions requires knowing the intended target object and how we want to convey its location to the listener. The bottom-up approach in image captioning is focused on learning the salience of objects and events to generate captions expressed in the dataset (e.g. Xu et al. (2015)). The combination of bottom-up and top-down approaches for generating descriptions use modularisation in order to improve the generation of descriptions of different kind (e.g. You et al. (2016)). However, as we have seen in the preceding discussion, the generation of spatial descriptions requires a highly specific geometric knowledge. How is this knowledge approximated by the bottom-up models? To what degree can we integrate this knowledge with the top-down models? In this paper, we investigate these questions in a language generation task by comparing different variations of included top-down spatial knowledge. More specifically, for each image, we generate a description for every pair of objects that are localised in the image. We consider a variety of top-down spatial knowledge representations about objects as inputs to the model: (a) explicit object localisation and extraction of visual features; (b) explicit identification of the target-landmark by specifying their order in the feature vector; and (c) explicit geometric representation of objects in a 2D image. We investigate the contribution of each of these sets of features to generation of image descriptions.

3 Neural Network Design

Our method is to add step-by-step modules and configurations to the network providing different kind of top-down knowledge in Section 2 and investigating the performance of such configura-

[2]Sometimes these are also known as *referent* and *relatum* (Miller and Johnson-Laird, 1976), *figure* and *ground* (Talmy, 1983) or *the located object* and *the reference object* (Herskovits, 1986; Gapp, 1994; Dobnik, 2009).

tions. There are several design choices with small effects on the performance but costly in terms of parameter size (Tanti et al., 2018b). Therefore, if there is no research question related to that choice, we take the simplest choice as reported in the previous work such as (Lu et al., 2017; Anderson et al., 2018). We use the following configurations:

1. Simple bottom-up encoder-decoder;
2. Bottom-up object localisation with attention;
3. Top-down object annotated localisation;
4. Top-down target and landmark assignment;
5. Two methods of top-down representation of geometric features (s-features).

These five configurations give us 10 variations of the model design as shown in Table 1. A detailed definition of each module is given in the Appendix A in the supplementary material.

Generative language model We use a simple forward recurrent neural model with cross-entropy loss in all model configurations.

Simple encoder-decoder An encoder-decoder architecture without spatial attention shown in Figure 3a and similar to (Vinyals et al., 2015) is the simplest baseline for fusing vision and language. The input to the model is an image and the start symbol $< s >$ of a description and the output is produced by the language model decoder. The embeddings are randomly initialised and learned as a parameter set of the model. The visual vectors are produced by a pre-trained ResNet50 (He et al., 2016). A multi-layer perceptron module (F_v in Figure 2) is used to fine-tune the visual features.

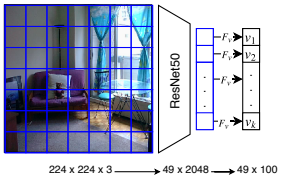


Figure 2: Visual features are obtained from the pre-trained ResNet50, then translated to a low dimensional vector with a dense layer F_v.

Bottom-up localisation With visual feature representing all regions of the image as in Figure 2, the attention mechanism is used as a localisation module. We generalised the adaptive attention introduced in (Lu et al., 2017) to be able to fuse the modalities. As shown in Figure 3b, the interaction between the attention mechanism and the language model is more similar to (Anderson et al., 2018): two layers of stacked LSTM, the first stack (LSTM_a) to produce the features for the attention model and the second stack (LSTM_l) to produce contextualised linguistic features which are fused with the attended visual features. This design is easier to extend with additional top-down vectors.

Top-down localisation Unlike the bottom-up unsupervised localisation, the top-down method includes a provision of a list regions of interest (ROI) from external procedures. For example, the region proposals can come from another bottom-up task as in (Anderson et al., 2018; Johnson et al., 2016) which use a Faster R-CNN (Ren et al., 2015) to extract possible regions of interest from the ConvNets regions in Figure 2. Here, as shown in Figure 4 we use the bounding box annotations of objects in images as the top-down localisation knowledge and then extract ResNet50 visual features from these regions. In the first stage the top-down visual representation only proposes visual vectors of the two objects in a random order without their spatial role as targets and landmarks in the descriptions. The model is shown in Figure 3d.

Top-down target-landmark assignment In the second iteration of the top-down localisation module we assign semantic roles to regions as targets and landmarks. This is directly related to localisation as spatial relations are asymmetric. We encode this top-down knowledge by fixing the order of the regions in the feature vector. The first object is the target and the second object is the landmark. Otherwise, the model is the same as in the previous iteration shown in Figure 3d.

Top-down geometric features The localisation procedure of objects discussed previously does not provide any geometric information about the relation between the two regions. However, top-down geometric features are required for grounding spatial relations where the location of the target object is expressed relative to the landmark. For example, a simple (but by no means sufficient) geometric relation between two bounding boxes can be represented by an arrow from the centre of one bounding box to the centre of the other and by ordering the information about bounding boxes in the feature vector as in the previous model to encode target-landmark asymmetry. The network architecture of the model with top-down geometric features expressing relations between the objects is shown in Figure 3e. We consider two different rep-

Model name	Regions Of Interest	TARGET-LANDMARK	s-features	Architecture
simple	-	-	-	Figure 3a
*bu*49	Bottom-up (7×7 grid)	Bottom-up attention	-	Figure 3b
*bu*49 + *mask*	Bottom-up (7×7 grid)	Bottom-up attention	Multi-hot 98	Figure 3c
*bu*49 + *VisKE*	Bottom-up (7×7 grid)	Bottom-up attention	Dense 11	Figure 3c
td	Top-down (2 bbox)	Bottom-up attention	-	Figure 3d
td + *mask*	Top-down (2 bbox)	Bottom-up attention	Multi-hot 98	Figure 3e
td + *VisKE*	Top-down (2 bbox)	Bottom-up attention	Dense 11	Figure 3e
td order	Top-down (2 bbox)	Top-down assignment	-	Figure 3d
td order + *mask*	Top-down (2 bbox)	Top-down assignment	Multi-hot 98	Figure 3e
td order + *VisKE*	Top-down (2 bbox)	Top-down assignment	Dense 11	Figure 3e

Table 1: The 10 variations of the neural network model after incrementally adding modules and features.

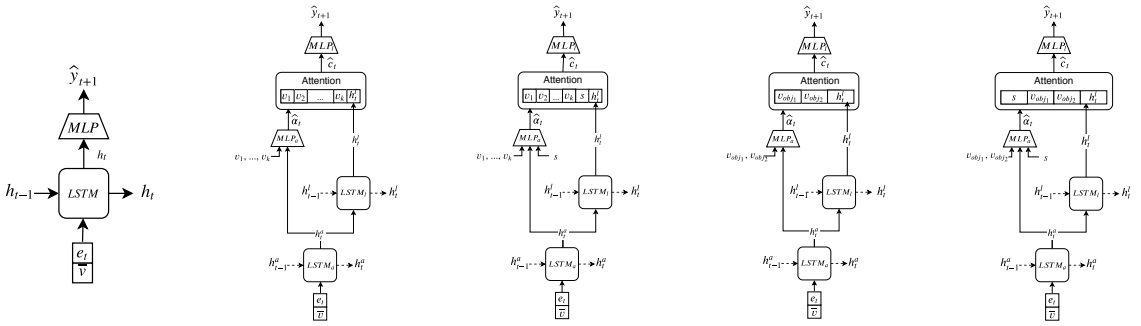


Figure 3: Five architectures: (a) simple encoder-decoder (*simple*). (b) bottom-up localisation with adaptive attention on 49 regions (*bu*49). (c) bottom-up localisation with explicit spatial vectors of the bounding boxes *bu*49 + *mask*/*bu*49 + *VisKE*. (d) top-down localisation with attentions on two bounding boxes (*td*). (e) top-down localisation augmented with explicit spatial vectors of the bounding boxes (*td* + *mask*/*td* + *VisKE*).

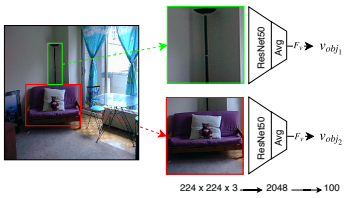


Figure 4: Top-down localisation of objects with bounding boxes whose visual features are extracted and translated to lower dimensions with F_v.

resentations of the top-down geometric features shown in Figure 5: Multi-hot mask over 49 vectors independently for target and landmark (*Mask*) over 49 locations (Figure 5a) and *VisKE* (Sadeghi et al., 2015) dense representations with 11 geometric features (Figure 5b) where dx, dy are changes in the coordinates of the centres, ov, ov_1, ov_2 the overlapping areas (total, relative to the first, and the second bounding box), h_1, h_2 heights, w_1, w_2 widths and a_1, a_2 areas. Note that *Mask* features provide geometric information about the size and the location of objects relative to the picture frame and *VisKE* feature provide more detailed geometric information that expresses the relation between the objects. The latter therefore more closely match the features that were identified in spatial cognitive models. A feed-forward network with two layers (F_s) is used to project geometric features into a vector with the same dimensionality as the F_v outputs so that different modalities are comparable in weighted sum model of attention.

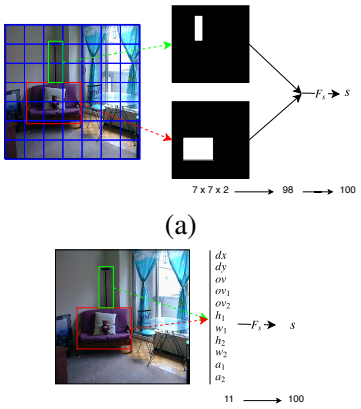


Figure 5: (a) Each bounding box is converted to a mask of multi-hot vector on 49 regions. (b) The geometric relation between the two bounding boxes are represented with features from (Sadeghi et al., 2015).

4 Dataset and Training

We use the relationship dataset in Visual Genome (Krishna et al., 2017) which is a collection of referring expressions represented as triplets

⟨subject, predicate, object⟩ on 108K images. Unlike image captioning datasets such as MSCOCO (Chen et al., 2015) and Flickr30K (Plummer et al., 2015) where only 5 captions are given for each image, each image in this dataset is annotated with 50 phrases. The annotators were asked to annotate relations between two given bounding boxes of subject and object by freely writing the text for each of the three parts of the annotation. The bounding boxes produced by another annotation procedure which detected objects in the images. In total, there are 2,316,104 annotations of 664,805 unique triplets, 35,744 unique labels of subjects and 21,299 unique labels of objects most of which consist of multiple tokens. We omit all repetitions of triplets on each image, this leaves total 1,614,055 annotations.[3]

Spatial relations Based on the lists of spatial prepositions in (Landau, 1996) and (Herskovits, 1986), we have created a dictionary of spatial relations and their possible multi-word variants including their composite forms. This dictionary contains 7,122 entries of 235 relations (e.g. *right* to represent both *on the right hand side of* and *to the right of*). Of these only 202 are found in Visual Genome dataset covering 79 spatial relations. 328,966 unique triplets in Visual Genome are based on exactly one of these terms which covers 49.4% of all possible relationships.[4]

Bounding boxes Each bounding box is a tuple of 4 numbers (x, y, w, h). We normalise the numbers to the range of $(0, 1)$ relative to the image size to create geometric feature vectors (Section 3). The image is split into a grid with 7x7 cells to which bounding boxes are mapped, one bounding box potentially covering more than one cell. With this bounding box granularity, there are exactly 308,330 possible bounding boxes. However, only 151,974 are observed in the relationships dataset. The spatial distribution of paired objects reflects how natural pictures are framed and how related objects are understood by annotators.

Pre-processing We first removed duplicate triplets describing the same image. Then we converted each triplet into a word sequence by concatenating the strings and de-tokenising them with the white space separator. This produced a corpus with a vocabulary of 26,530 types with a maximum sequence length of 16 tokens and on average 15 referring expressions per image. We use 95% of the descriptions for training and 5% for validation and testing (5,230 images with 80,231 triplets).

Training We use Keras (Chollet et al., 2015) with TensorFlow backend (Abadi et al., 2015) to implement and train all of the neural network architectures in Section 3. The models are trained with the Adam optimiser (Kingma and Ba, 2014) ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with a batch size of 128 and 15 epochs.

5 Evaluation

All implementations are available online[5].

5.1 Qualitative Examples

Figure 6 shows generated descriptions for two examples of unseen pictures from the test dataset by five models. The generated word sequence is that with the lowest loss using beam search with $k = 5$. The first example shows exactly how top-down localisation of objects is important especially if the goal is to refer to specific objects in the scene. In the second example, the visual features inside the bounding box are confusing for all 5 models. More examples are in Figure 13 in the Appendix.

5.2 Overall Model Performance

Hypothesis Top-down spatial knowledge improves the model performance. We consider three categories of top-down spatial knowledge: (i) top-down localisation of regions of interest; (ii) top-down assignment of semantic roles to regions; and (iii) two kinds of geometric feature vectors.

Method After training the models we evaluate them by calculating the average word level cross-entropy loss on held out instances in the test set[6].

[3] The repetitions include reflexive expressions (e.g. *horse next to horse*), annotations of several objects of the same type (e.g. *cup on table*), and repetitions due to several bounding box annotations of the same objects with different sizes.

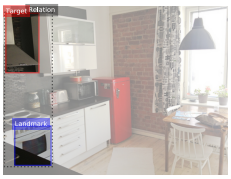
[4] Other triplets in Visual Genome also have spatial content. Some of them include modifiers such as *partially under* as in Figure 1 and some of them are descriptions of an event or an action such as *sitting on* and *jumping over*. Some annotated relationships are verbs such as *flying* with less obvious spatial denotation. The spatial bias in the dataset was studied in (Collell et al., 2018). The most frequent spatial relation in the dataset is "*on*" (over 450K instances), the second place is "*in*" (150K instances), then "*with*", variations of "*behind*", "*near*", "*top*", "*next*", "*under*", "*front*", and "*by*" (less than 10K instances each).

[5] https://gu-clasp.github.io/generate_spatial_descriptions/

[6] Equivalent to log-perplexity of the language model.

⟨ “bat”, “over”, “shoulder”⟩

simple	player
*bu*49	man wearing shirt
td	bat in hand
td order	bat in hand
td order + *VisKE*	bat in hand



⟨ “hood”, “above”, “oven”⟩

simple	window
*bu*49	pot on stove
td	oven has door
td order	vent above sink
td order + *VisKE*	cabinet has door

Figure 6: From VisualGenome: 2412051[a] 2413282[b]

[a] Herholz (2005): CC BY-SA 2.0.
[b] juanjogasp (2013): CC BY-NC-SA 2.0.

We also calculate the loss on descriptions containing specific spatial relations for qualitative understanding of the effects of each type of top-down knowledge.

Results The overall loss of each model on the unseen descriptions of images is shown in Figure 7. The fully bottom-up model with no spatial attention (*simple*) has the highest loss. The loss in the variations of the model with bottom-up localisation in *bu*49 is higher than the one in the models with top-down localisation. The models with the top-down assignment of TARGET-LANDMARK achieves the best results. The effect of top-down geometric features is not significant.

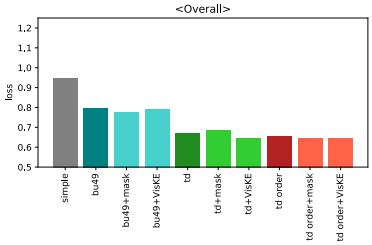


Figure 7: Cross-entropy loss of different model configurations on evaluation data.

Figure 8 shows the performance of the models on a selection spatial relations.

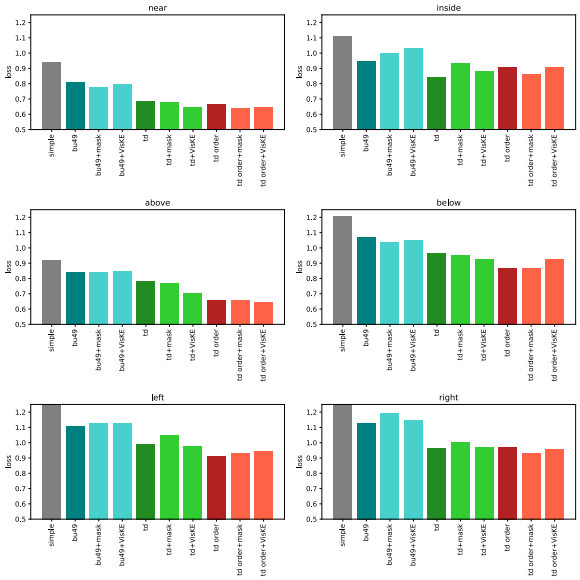


Figure 8: Cross-entropy loss of different model configurations on 40 descriptions for each relation: *near*, *inside*, *above* and *below*.

Discussion The top-down localisation (*td*) certainly improves the performance of the language models compared to purely bottom-up representations. However, additional top-down assignment of TARGET-LANDMARK (*td order*) and their additional geometric arrangement of bounding box features (*mask* and *VisKe*) has a small positive effect on overall performance. The overall performance is not a representative of how these configurations effect the grounding of spatial relations. More specifically, the imbalance of certain groups of relations (especially a generally lower proportion of geometrically biased relations such as “left” and “right” in this dataset and the presence of relations with a minimum spatial content such as *has*, *wearing*) makes it harder to make conclusions about overall performance of the models. We further examine two groups of some frequent spatial relations. The relations such as *inside* and *near* represent one group and *above* and *below* represent the other. Some top-down knowledge (as represented by our features) is less informative for the first group but is informative for the second group. For example *near* does not require the assignment of TARGET-LANDMARK roles. We observe that *td order* is not performing better than *td*. On the other hand, *inside* is sensitive to TARGET-LANDMARK assignment. However, since the relation is also restricted by a choice of objects (only certain objects can be inside others) their TARGET-LANDMARK assignment may already be

inferred without such top-down knowledge from a language model. For the second group, the top-down knowledge about the semantic role of objects is important. However, *left* and *right* are among the least frequent relations in the dataset which is demonstrated by the fact that their descriptions have a higher loss than *above* and *below*. For these relations the loss of the *simple* model is much higher than other configurations. It can be seen that *td* is performing better than *bu* and *td order* is contributing over *td* but geometric features have a lesser effect than identification of semantic roles (*td order*).

5.3 Grounding in features

Hypotheses With the aim to evaluate *what* top-down information contributed to grounding of words we examine the following hypotheses:

H1 s-features contribute to predicting spatial relation words.

H2 Without top-down TARGET-LANDMARK role assignments to each region, attention is uniformly distributed over region choices at the beginning of a sequence generation.

Method In order to check the contribution of each feature from different modalities in prediction of each word, we look at the adaptive attention on each feature at the point of predicting the word[7]. Since feature vectors are not normalised against the number of features of each modality, we first multiply each attention measure with the magnitude of the feature vector, and then we normalised it to sum to 1 again:

$$\beta_{t,f_i} = \frac{\alpha_{t,f_i}||f_i||}{\sum_j \alpha_{t,f_j}||f_j||} \quad (1)$$

where t refers to the time in the word sequence, and f_i is the feature the attention of which α_{t,f_i} is applied to it. We report the average β_{t,f_i} over the instances in the validation dataset.

Figure 9 shows β on two examples in three models. For each word, the bar chart is divided between four features (in Figure 3e): (1) target v_{obj_1} (2) landmark v_{obj_2} (3) s-features for bounding boxes (4) contextualized embeddings h^l.

[7] In this experiment, we do not check if the estimated likelihood for the correct word is the highest predicted score. The generated descriptions may still be acceptable with an alternative spatial relation. Furthermore, in the following analysis we report the attention over semantic roles and not individual words.

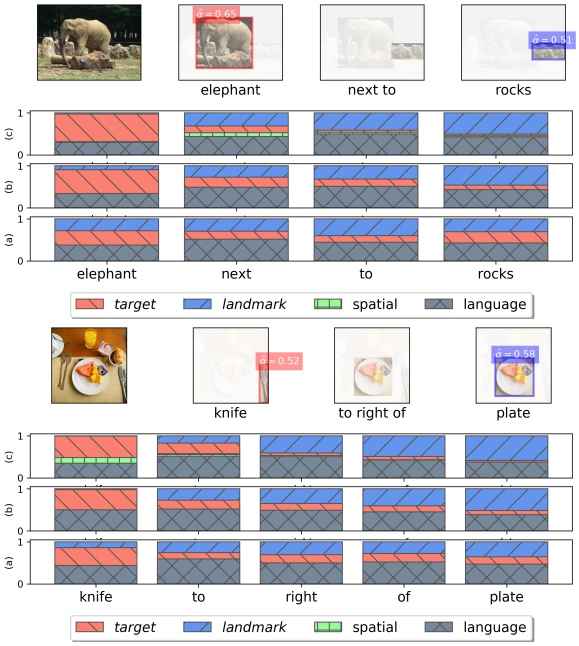


Figure 9: β is plotted in bar charts for each word. (a) *td order* + *VisKE* (b) *td* + *VisKE* (c) *td*. The values of β for each word that constitute description referring to each bounding box region is given in images.

After measuring the normalised attention on each feature according to Equation 1, we report the average of attentions on each token at that time step of the word sequence. We also group the tokens based on their semantic role in the triplets and report the average β on these tokens for a given role.

Results The average of attentions over triplets of tokens is plotted in Figure 10. The behaviour of attentions on word sequences in the four models in given in Figure 11.

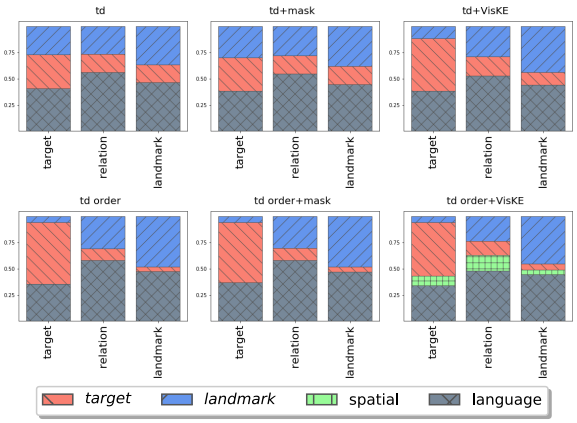
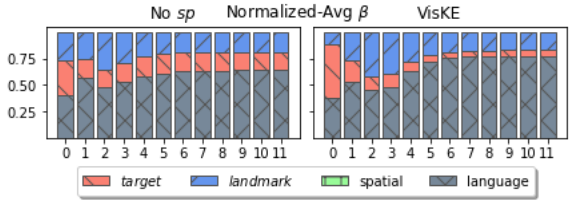
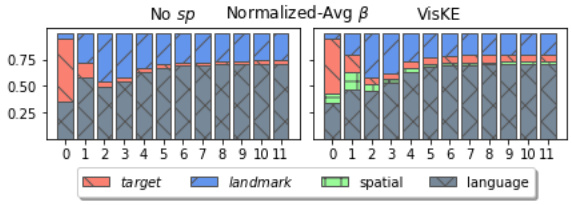


Figure 10: The overall average of β on tokens of each semantic role (target, relation, landmark) on all examples of the test dataset, for 6 variations of the top-down knowledge about regions of interest (ROI): location of objects and their order as target and landmark.



(a)



(b)

Figure 11: The average of β attentions of top-down models over sequences of words 1...11 (a) comparing *td* and *td* + *VisKE* and (b) comparing *td order* and *td order* + *VisKE*.

Discussion The comparison of 6 models in Figure 10 shows that geometric *mask* s-features are not contributing as well as dense *VisKE* s-features. In the models without top-down semantic role assignment only the model with +*VisKE* features has the expected attention on target and landmark, but there is no attention on the s-features. In the models with top-down semantic role assignment, the model with *VisKE* s-features has higher attention on s-features when predicting a relation word (H1). A similar situation is observable over word sequences in Figure 11. Without prior semantic role assignment the model is more confused how to attend target or landmark (H2). Finally, note that geometric *VisKE* s-features help predicting the TARGET-LANDMARK roles when these are not assigned top-down.

6 Related Work

Generating referring expressions Generating locative expressions is part of the general field of generating referring expressions (Dale and Reiter, 1995; Krahmer and van Deemter, 2011) with applications such as describing scenes (Viethen and Dale, 2008) and images (Mitchell et al., 2012). The research on describing visible objects (Mitchell et al., 2013) and human-robot dialogue (Kelleher and Kruijff, 2006) raised question about grounding relations in hierarchical representation of context. Application of neural language models and using convolutional neural networks for encoding visual features is an open question in interactive GRE tasks.

Encoder-decoder models with attention Recently several methods focused on finding better neural architectures for generating image descriptions based on pre-trained convolutional neural networks have been introduced. Karpathy and Fei-Fei (2015) align descriptions with images. Vinyals et al. (2015) introduce an encoder-decoder framework. Xu et al. (2015) improve this approach with spatial attention. Lu et al. (2017) introduce adaptive attention that balances language and visual embeddings. The attention measure provides an explanation of encoder-decoder architectures on how each modality contributes to language generation. Based on the attended features the performance of these models can be examined (Liu et al., 2017; Ghanimifard and Dobnik, 2018). In our paper, we develop a model similar to the adaptive attention which exploits its expressive aspects as a degree of grounding in different features.

Outputs of external models as top-down features In another line of work, the output of the bottom-up visual understanding is used as top-down features for language generation. For example, an object detection pipeline is combined explicitly with language generation. This procedure was previously used in template-based language generation (Elliott and Keller, 2013; Elliott and de Vries, 2015). There have been attempts to combine this process with neural language models with attention. For example, You et al. (2016) extract candidate semantic attributes from images (e.g. a list of objects in the scene), then the attention mechanism is used to learn to attend on them when generating tokens of image descriptions. Instead of semantic attributes, Anderson et al. (2018) use a region proposal network from a pre-trained object detection model to extract the generated bounding box regions as possible locations of visual clues. Then, the attention model learns to attend on the visual features associated with these regions. The idea of using an object detection module is also used in Johnson et al. (2016) where Faster R-CNN (Ren et al., 2015) is used to find regions of interest. Instead of assigning one object class to each region, a full description is generated for each proposed region. In all of these models, an image understanding module extracts some proposed representations and then this knowledge is used as a top-down representation of the scene to generate an image description. In this paper, we investigate the extent to which differ-

ent spatial information is facilitating as a top-down knowledge to generate descriptions of scenes with neural language models.

Modular design Our paper examines strategies that can demonstrate language grounding within a neural architecture. The studies of neural architectures such as (Tanti et al., 2018b) provide analytical insight on differences between multimodal architectures for language generation. The modular design is mostly used in language parsing tasks such as (Hu et al., 2017) where object recognition, localisation and relation recognition are separate modules for grounding different parts of image descriptions in images in order to solve tasks such as visual question answering. In our paper, the modularity of the neural architecture is not focused on parsing text but used to incrementally demonstrate the contribution of each introduced modality to language generation.

Multimodal embeddings There are related studies on learning multimodal embeddings (Kiros et al., 2014; Lazaridou et al., 2015) to represent vision and language in the same semantic space. The focus of our paper is to investigate how these different modalities complement each other in neural language generation. In our models, the semantic representations of spatial relations are considered as a separate modality extending both the language and visual embeddings. There are related studies on encoding spatial knowledge in feature space in order to predict spatial prepositions (Ramisa et al., 2015) or on prepositional embeddings which can predict regions in space (Collell and Moens, 2018). In our paper, we investigate the degree in which each embedding contributes to language generation within the neural language model.

7 Conclusions

We explored the effects of encoding top-down spatial knowledge in a bottom-up trained generative neural language model for the image description task. The findings of the experiments in this paper are as follows:

(1) Overall, integration of top-down knowledge has a positive effect on grounded neural language models for this task. (2) When combining bottom-up language grounding with top-down knowledge representation as different features, different types of top-down knowledge have different contribution to grounded language models. The general picture is further complicated by the fact that different spatial relations have different bias to different knowledge. (3) The performance gain from the geometric features extracted from bounding boxes (s-features) is smaller than initially expected, with two possible explanations related to the nature of the corpora of image descriptions: (i) The corpus contains images of typical scenes where the relation of objects with each other is predictable from the description and therefore is captured in the language model; (ii) As annotators are focused on describing "what is in the image" rather "where things are spatially in relation to each other", descriptions of geometric spatial relations which refer to the locational information are rare in the corpus. (4) The majority of attention is placed on the language model which demonstrates that this provides significant information when generating spatial descriptions. While this may be a confounding factor if the visual features are ignored, the language model also encodes useful information about spatial information as discussed in (Kulkarni et al., 2011; Dobnik et al., 2018).

The results open several questions about grounded language models. Firstly, the degree to which the system is using each modality can be affected by dataset biases and this should be taken into account in the forthcoming work. Given this bias, learning a single common language model for descriptions of spatial scenes is insufficient as different kinds of knowledge may come to focus in different interactional scenarios. This further supports the idea that top-down integration of knowledge is required where we hope that the models will learn to attend to the appropriate features. Secondly, our investigation leaves open the question whether the representations both visual and geometric that we use are good representations for learning spatial relations. Further work will include a focused investigation of what kind of geometric relations they encode.

Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *CVPR*, 3(5):6.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

François Chollet et al. 2015. Keras. `https://keras.io`.

Michael Coghlan. 2011. Tony cook. VisualGenome image id 2413371.

Guillem Collell and Marie-Francine Moens. 2018. Learning representations specialized in spatial knowledge: Leveraging language and vision. *Transactions of the Association of Computational Linguistics*, 6:133–144.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kenny R Coventry, Merce Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of memory and language*, 44(3):376–398.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302.

Desmond Elliott and Arjen de Vries. 2015. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 42–52.

Klaus-Peter Gapp. 1994. Basic meanings of spatial relations: computation and evaluation in 3D space. In *Proceedings of the twelfth national conference on Artificial Intelligence (AAAI'94)*, volume 2, pages 1393–1398, Menlo Park, CA, USA. American Association for Artificial Intelligence, AAAI Press/MIT Press.

Mehdi Ghanimifard and Simon Dobnik. 2018. Knowing when to look for what and where: Evaluating generation of spatial descriptions with adaptive attention. In *Proceedings of the 1st Workshop on Shortcomings in Vision and Language (SiVL'18), ECCV, 2018*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Dave Herholz. 2005. Wide stance. VisualGenome image id 2412051.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling

relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1124.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

juanjogasp. 2013. Baltic trip. VisualGenome image id 2413282.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12 –13 June*, volume 1 of *CLASP Papers in Computational Linguistics*, pages 41–52, Gothenburg, Sweden. Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, CLASP, Centre for Language and Studies in Probability.

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1041–1048, Sydney, Australia. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603.

Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

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.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.

Barbara Landau. 1996. Multiple geometric representations of objects in languages and language learners. *Language and space*, pages 317–363.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017. Attention correctness in neural image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Gordon D Logan. 1994. Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5):1015.

Gordon D Logan. 1995. Linguistic and conceptual control of visual spatial attention. *Cognitive psychology*, 28(2):103–174.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, University of Oxford: Committee for Comparative Philology and General Linguistics, Oxford, United Kingdom.

George A. Miller and Philip N. Johnson-Laird. 1976. *Language and perception*. Cambridge University Press, Cambridge.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1184, Atlanta, Georgia. Association for Computational Linguistics.

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 *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Arnau Ramisa, JK Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220. Association for Computational Linguistics.

RaSeLaSeD_Il_Pinguino. 2008. Killer bear. VisualGenome image id 2318741.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

Jared Schmidt. 2010. Desk 2010-08-08. VisualGenome image id 2413204.

Leonard Talmy. 1983. How language structures space. In Herbert L. Pick Jr. and Linda P. Acredolo, editors, *Spatial orientation: theory, research, and application*, pages 225–282. Plenum Press, New York.

Marc Tanti, Albert Gatt, and Kenneth P Camilleri. 2018a. Quantifying the amount of visual information used by neural caption generators. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0.

Marc Tanti, Albert Gatt, and Kenneth P Camilleri. 2018b. Where to put the image in an image caption generator. *Natural Language Engineering*, 24(3):467–489.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Brian Yap. 2008. New england highway. VisualGenome image id 2417890.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.