Abstract. The task of automatically describing an image with natural language requires techniques to associate visual representations with their corresponding linguistic units. In the state of the art techniques, most commonly, a pre-trained convolutional neural networks extracts visual features of an image, and then a neural language model with attention mechanism is trained as a decoder to generate descriptions. However, as argued in the previous work such networks are good at detecting objects but are not using and describing geometric relations between objects. In this paper we explore the possibility of using the location of objects as explicit feature in generating image captions. We improve the state of the art image captioning model with adaptive attention described in [13] with explicit object locations derived from the annotated bounding boxes and investigate to what degree each kind of visual representations is contributing to prediction of spatial relations between objects: (i) visual object features from a pre-trained CNN and (ii) explicit object locations.

Keywords: spatial relation recognition · object recognition · image descriptions · neural language model · grounded language model

1 Introduction

Automatically generating descriptions of scenes involving objects in images is an interdisciplinary research problem. A common framework using the state-of-the-art neural network tools is an attention-based encoder-decoder model as implemented in [17,18,13,1]. In these models the attention mechanism is used to align visual clues with linguistic units or words. The visual features are often obtained from convolutional neural networks that were pre-trained for object recognition task on ImageNet [2]. The spatial attention works as a detection module which aligns visual features in regions of an image with words and so ensures that their generation is based on visual clues.

In the caption generation task the spatial attention on the output of the convolutional neural networks was initially suggested as a method to identify salient objects [17]. [13] improved this method by providing the possibility of
balancing between the visual features and the features from the recurrent language model. The application of attention modules is not limited to this concept of spatial attention. For example, [18] lists recognisable attributes in an image, and then an attention mechanism in a recurrent neural language model is applied on them to generate a description. [9, 4] argue and demonstrate that in a grounded neural language model the generation of spatial relations such as “to the left of” in spatial descriptions such as “the chair is to the left of the table” is less dependent on visual features compared to noun phrases due to the fact that the learned visual features are suited for object recognition and not recognition of geometric spatial relations between objects. It follows that in image captioning systems based on DNNs features representing such geometric relations should be incorporated. There are neural network designs such as [8] that use explicit object detection modules where the locations of objects are given to the model to extract their features. In this paper we explore the idea of using spatial information as features for generating spatial relations in natural language descriptions.

Spatial descriptions are composed of a target “the chair”, a landmark “the table” and a spatial relation between them “to the left of”. The target and the landmark are noun phrases while the spatial relation is normally a prepositional phrase. The previously mentioned DNN image captioning models face with the following difficulties: (i) not all linguistic units are associated with visual clues in an image – as a solution adaptive attention [13] was proposed which can attend either on the visual features or the features of a recurrent language model; (ii) the sub-sampling in convolutional neural networks fails to preserve the geometric spatial relation between high level parts in a way that would represent geometric spatial relations [6, 9]; (iii) generating of spatial relations should not be based on features in one specific region of an image. Instead, the features must represent relational information between objects rather the localised visual clues.

In this paper we explore the contribution of visual and geometric features in generating spatial image descriptions. Our expectation is that adding information that describes geometric relations between the objects will improve the language model generating such descriptions. In order to simplify the task to directly observe the effect of the added object location features, we do not rely on the predicted spatial attention over objects for their location information but this is obtained directly from the annotated bounding boxes. The idea that two mechanisms are involved in the apprehension of spatial scenes, one focused on detecting objects and one on detecting geometric relational information between them, is also confirmed by the studies in spatial cognition and language development. [11] argues that descriptions of objects and descriptions of locations draw on different kinds of spatial representations (“what” and “where”) and that these differences may be correlated with a property of neurological design. In our work, due to the way DNN models work, we expect that different descriptions will be generated on the basis of information from each source to a different degree.
2 Method

A simple encoder-decoder architecture without spatial attention, similar to [16], provides a base system for our tests of interaction of visual and geometric features in generating spatial descriptions in an end-to-end architecture. It has been shown that such an architecture on overall achieves a lower performance than an architecture with spatial attention [17]. The input to the model is an image and the start symbol $<s>$ to the language model which then predicts the following word. The word embeddings ($e_t$ in Figure 1a) are randomly initialised and concatenated with the average of visual features ($\bar{v}$) which define the general visual representation of the scene. A pre-trained ResNet50 [5] is used for extracting the visual features. The visual vectors are translated to a lower unified vector size for computational convenience through a fine-tuning layer. The resulting concatenated vector containing a word embedding and a visual vector is fed to a Long-Short Term Memory (LSTM) network [7] which learns a grounded language model. The output of the LSTM is fed to a multi-layer perceptron (MLP) with a softmax layer which predicts the next word:

$$P(w_{t+1} = \hat{y}_{t+1}|V, w_{1:t}) = \text{Softmax}(\text{MLP}(\text{LSTM}(e_t; \bar{v}, h_{t-1})))$$ (1)

where $e_t$ represents the word embedding in time $t$, $h_{t-1}$ represents a hidden unit in recurrent cell, $V = \{v_1, ..., v_k\}$ corresponds to the visual features and $\bar{v}$ is the average of the visual vectors.

Inspired by [13], we use the adaptive attention to balance the contribution of visual features and linguistic features. This model learns a degree to attend each feature type when predicting a word (Figure 2a). However, as our earlier work shows [4], the attention mechanism learns how to associate words describing objects and their visual locations but words describing spatial relations are less attended visually and also the attentional mask over the scene does not correspond well with what we would consider a good spatial template for such a relation [12]. The question we examine here, therefore, is whether spatial relations would
be attended better in perceptual features by the attention mechanism if these features would be more closely aligned with the features used in spatial cognition. In the second model (Figure 2b) we use the bounding box information to represent the visual regions occupied by the objects. The model in this case will be trained to attend the visual representations within the two bounding boxes locating the objects. In the third variant of the model (Figure 2c), in addition to the visual features within the two bounding boxes, geometric information extracted from the bounding boxes is also explicitly added. In (b) and (c) the bounding boxes and the geometric features are represented as vectors, then a feed-forward network with two layers ($F_s$) is trained to project them into a 100-dimension vector.

$$F_s : W_s^2 \in \mathbb{R}^{100 \times 100}, W_s^1 \in \mathbb{R}^{100 \times 11} (or \mathbb{R}^{100 \times 98})$$

$$s = W_s^2 tanh(W_s^1 s' + b_s)$$

Two bounding box vectorisation strategies used in the test are shown in Figure 2. (1) The simple vectorisation concatenates two mask vectors (of the length of 49) to produce 98 dimensions (Figure 2b). (2) A dense vectorisation which represents the bounding boxes as 11 geometric features: where $dx, dy$ are distances between 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 according to [15] (Figure 2c).

The multi-layer perceptron for attention model is formalised as follows:

$$z = W^2 a tanh(W^1 a V + W^1 h^a_T \mathbb{1}^T) \quad (2)$$

$$\hat{\alpha}_t = \text{Softmax}(z) \quad (3)$$

where the dimension of $W^2 a$ is chosen based on the number of feature vectors to be attended. The attention $\hat{\alpha}_t$ pools the features from the given list of vectors:

$$c_t = \sum_{i=1}^{k} \alpha_{t,i} v_i + \alpha_{t,k+1} s + \alpha_{t,k+1} h_i^t. \quad (4)$$

### 3 Evaluation and results

We train 5 models (a simple model, a model using adaptive spatial attention, a model using visual features within the bounding boxes of objects, a model with a mask over the bounding boxes of objects, and a model using visual features within the bounding boxes of objects and geometric features of bounding boxes) on 1.6 million samples from Visual Genome [10]. The dataset contains 108K images with several annotations per image. In order to maximise the the number of spatial descriptions, we use the relationship dataset to generate referring expressions which contains annotations of target, landmark and relationship triplets. First we concatenate the triplets (each may contain several words) into a locative phrase, then we tokenise this phrase into a sequence of individual words.
Fig. 2. Three different models of attention: (a) adaptive spatial attention similar to [13], (b) attention over visual vectors within the two bounding boxes, and (c) attention over object visual vectors as in (b) and explicit geometric spatial vectors.

Fig. 3. (a) The bounding boxes crop original images and then ResNet50 with average pooling extracts the visual features from these areas. (b) Each image is converted to a mask vector with 49 regions in which the two bounding boxes are represented. (c) The geometric properties of the two bounding boxes are represented by a vector of 11 features from [15].
Overall performance During training each model converged with a different level of cross-entropy loss (Figure 4). The three models with bounding boxes identifying the two objects have the best performing results but their performance is very similar. Adding representations of the bounding boxes as a feature vector in both cases improves the results but the improvement is not noticeable.

![Cross-entropy loss on validation set after 5 epochs over 1.6 million samples.](image)

**Fig. 4.** Cross-entropy loss on validation set after 5 epochs over 1.6 million samples.

Grounding in features A qualitative inspection shows that objects are correctly attended in their corresponding bounding boxes. In case of the model with explicit geometric representations, generating spatial relation terms puts higher attention on the spatial feature vector (Figure 5).

4 Conclusion

We explored the possibility of representing geometric properties of object bounding boxes as feature vectors in an end-to-end image description system. Our results indicate that grounding spatial relations in such representations is possible but the general level of improvement is not significantly higher than the case when geometric features are not introduced explicitly. We also note that 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 [3].

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. Secondly, our investigation still leaves open the question whether the representations both
Visual grounding of spatial relations in recurrent neural language models

Fig. 5. Three examples of attention behaviour: (1) attention on the language model ($h_t$), (2) attention on the spatial features of bounding boxes ($s$), (3) attention on the visual features in the first bounding box ($obj_1$, the target) and (4) attention on the visual features in the second bounding box ($obj_2$, the landmark). The bold numbers are the highest values either in each line or each column.
visual and geometric that we use are good representations for learning spatial relations. In our future work we plan (i) a more thorough investigation of the CNNs and their ability of encoding information useful for predicting spatial relations; and (ii) a more thorough investigation of the contribution of the geometric features. For (i), although CNNs do not encode precise geometric relations between objects, they still may encode useful discriminatory visual features, for example connectedness of particular features identifying configurations of object, for spatial relations. For (ii) the bounding boxes and the VisKE features are still quite primitive geometric representations that may not be exploited by the model. One extension is to use a more sophisticated model of spatial features based on psychological experiments [14]. In Figure 5 in all examples the bounding boxes between the objects overlap when the bounding boxes are projected relative to the image frame rather than the landmark. Consequently, all geometric representations are very similar and therefore the language model may be far more discriminative feature to predict a relation.

References