Visual grounding of spatial relations in recurrent neural language models

Mehdi Ghanimifard and Simon Dobnik

Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
firstname.lastname@gu.se

Abstract. The task of automatically describing an image with natural language requires techniques to associate linguistic units with their corresponding visual representations. In the state of the art techniques, most commonly, a pre-trained convolutional neural networks extracts visual features of the image, then a neural language model with attention mechanism will be trained as a decoder to generate descriptions. In this project, we explore the possibility of using the location of objects as explicit features to detect spatial relations between them in the recurrent neural language model.

Our attempt is to improve the state of the art models but the more important question we ask is in what order each visual representations are contributing for predicting spatial relations between objects: 1. object recognition features from pre-trained network. 2. explicit object locations. We extend the adaptive attention model in [10] with some modifications. Including the simplification of the task with bounding box information and injecting explicit object location encodings. Then we report

Keywords: Spatial recognition · Object recognition · Image description · Neural Language Model · Grounded language model

1 Introduction

Automatically generating description for scenes and objects in images is an interdisciplinary research problem. A common framework which benefits from the state of the art neural network tools is attention-based encoder-decoder model [13, 14, 10, 1]. In these models, the attention mechanism is designed to align visual clues with linguistic units. Ofter, the visual features are obtained by a convolutional neural networks with pre-trained weights for object recognition task in ImageNet [2]. Practically, the spatial attention is a detection module, which aligns visual features in regions of image with words which their generation would be based on these visual clues. We argue with current visual representations, this method is not enough for language grounding.

The spatial attention on convolutional neural networks initially suggested for caption generation task as a method to find salient objects [13]. Similarly,
Mehdi Ghanimifard and Simon Dobnik

[10] improved this method by providing the possibility of balancing between the relevant features. The application of attention modules is not limited to spatial attention, for example [14] lists recognisable attributes in an image, then an attention mechanism in recurrent neural language model is applied on these attributes to generate the description. The neural network designs with explicit object detection modules such as [6], the locations of objects given to the model only to extract features. In this project we explore the idea of using spatial information as features for generating spatial relations in natural language.

When generating spatial expressions, the composed sentence has a nominal phrase for target, a nominal phrase for the landmark and a spatial relation between them. However, the models mentioned above come with three important complications: (1) Not all linguistic units are linked to specific visual clues in image. As a solution an adaptive attention like [10] can attend on two different features. (2) The subsampling in convolutional neural networks fail to preserve the spatial relation between high level parts [4, 7]. (3) Generating spatial relations shouldn’t be based on features in one specific region of image. Instead, the features must represent relational features regions rather the local visual clues.

In this project, we explore the effect of each feature in language model for image descriptions. The expectation is that adding additional information must improve the language model. In order to simplify the task in favor of observing the direct effect of added features, we simplify the spatial attentions to object selection. The significance of this question can also be discussed in the framework of previous studies in spatial cognition and language development. [9] argues that although the object recognition and spatial recognition are related, the difference between the number of object names and the number of spatial terms in the English language might be an evidence for different cognitive representations for “what” and “where”. In this work, we consider this to be a mixture of two representation in different degrees.

2 Method

A simple encoder-decoder architecture without spatial attention, similar to [12], has been proven to have lower performance comparing to the cases with spatial attention. However it provides a baseline to understand the effect of each module in an end-to-end architecture (Figure 1).

The input to the model is an image and the start signal \( s_i \) to the language model for predicting the next word. The words embeddings in this setting are randomly initialized and concatenated with the average of visual features as a general visual representation of the scene. In the implementation of the a pretrained ResNet50 [3] for extracting visual features is used. Which the visual vectors got translated to lower unified vector size for computational convenience. The word embedding and visual vector combined with a Long-short term memory (LSTM) [5] make the language model. The output of the LSTM is fed to another multi layer perceptron (MLP) with softmax layer for predicting the next word:
Visual grounding of spatial relations in recurrent neural language models

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

(1)

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

\[
F_v : W_v \in \mathbb{R}^{100 \times 2048}, b_v \in \mathbb{R}^{100} \\
v_i = \text{ReLU}(W_v v'_i + b_v) \\
\bar{v} = \sum_{i=1}^{k} v_i
\]

 Inspired by [10], we use the adaptive attention between visual features and linguistic features. This model learns in what portion attend to each type of feature (Figure 2).

Fig. 2. Three different models: (a) spatial attention with adaptive attention (similar to [10]) (b) simplified visual vectors only with two visual vectors for two objects. (c) object vectors similar to (b) with additional explicit representation of spatial relations.

In order to examine the effect of spatial attention, we use bounding box information to generate more simplified representation of regions. The model in this case will be trained to attend to two given representations referring to
two objects. In third modification, in addition to object vectors, the bounding box information are used as explicit representation of relational information. The bounding box information are vectorized with two different strategies (Figure 4). The details of spatial attention is provided in Appendix 5.

3 Evaluation and Results

We train the models on 1.6 million samples from Visual Genome [8]. The dataset has 108K images with several annotations per image. In order to maximise the exposure to spatial relations, we used the relationship dataset to generate referring expressions. First we concatenate the triplets as a phrase, then we tokenised it to get the word sequence.

Language model After training the model, each converged with different level of cross-entropy loss (Figure 3). Three models with two objects have the best performing results but very similar performance. Adding the representation of space as a feature vector in both cases improves the results but the improvement is noticeable.

![Fig. 3. The cross-entropy loss on validation set, after 5 epochs over 1.6 million samples.](image)

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

4 Conclusion

We explored the possibility of representing spatial relations as feature vectors in an end-to-end image description system. Our qualitative observation indicates that grounding spatial relations in such representations is possible but the general level of improvement is not significantly higher than the case when the features are not introduced explicitly.
References


5 Appendix: Equations

The multilayer perceptron for attention model formalised as follows:
\[ z = W^2 \tanh(W^1 a_t V + W^1 a_t h_{t-1}^T) \]  
\[ \hat{\alpha}_t = \text{Softmax}(z) \]

Where the dimension of \( W^2 \) 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:

\[ c_t = \sum_{i=1}^{k} \alpha_{t,i} v_i + \alpha_{t,k+1}s + \alpha_{t,k+1}h_l^j \]

6 Appendix: Images

Fig. 4. (a) The spatial relation between two bounding box is represented with a vector of 11, inspired from [11]. (b) each bounding box will be converted to a mask vector for 49 regions, a concatenation of them represents two boxes. (c) the minimum square covering each bounding box corps images from the original source, then ResNet50 with average pooling captures the visual features of the image areas.
Visual grounding of spatial relations in recurrent neural language models

Fig. 5.

<table>
<thead>
<tr>
<th>$h_t$</th>
<th>$s$</th>
<th>obj1</th>
<th>obj2</th>
<th>word</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.796</td>
<td>0.054</td>
<td>0.111</td>
<td>0.039</td>
<td>topping</td>
</tr>
<tr>
<td>0.876</td>
<td>0.034</td>
<td>0.030</td>
<td>0.060</td>
<td>on</td>
</tr>
<tr>
<td>0.846</td>
<td>0.002</td>
<td>0.018</td>
<td>0.134</td>
<td>a</td>
</tr>
<tr>
<td>0.849</td>
<td>0.003</td>
<td>0.009</td>
<td>0.140</td>
<td>pizza</td>
</tr>
<tr>
<td>0.900</td>
<td>0.004</td>
<td>0.011</td>
<td>0.085</td>
<td>EOS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$h_t$</th>
<th>$s$</th>
<th>obj1</th>
<th>obj2</th>
<th>word</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.802</td>
<td>0.028</td>
<td>0.165</td>
<td>0.005</td>
<td>man</td>
</tr>
<tr>
<td>0.773</td>
<td>0.170</td>
<td>0.020</td>
<td>0.037</td>
<td>in</td>
</tr>
<tr>
<td>0.839</td>
<td>0.031</td>
<td>0.008</td>
<td>0.121</td>
<td>jacket</td>
</tr>
<tr>
<td>0.899</td>
<td>0.012</td>
<td>0.013</td>
<td>0.076</td>
<td>EOS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$h_t$</th>
<th>$s$</th>
<th>obj1</th>
<th>obj2</th>
<th>word</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.772</td>
<td>0.061</td>
<td>0.160</td>
<td>0.007</td>
<td>head</td>
</tr>
<tr>
<td>0.853</td>
<td>0.038</td>
<td>0.075</td>
<td>0.034</td>
<td>of</td>
</tr>
<tr>
<td>0.842</td>
<td>0.007</td>
<td>0.026</td>
<td>0.126</td>
<td>a</td>
</tr>
<tr>
<td>0.913</td>
<td>0.002</td>
<td>0.009</td>
<td>0.075</td>
<td>EOS</td>
</tr>
</tbody>
</table>