Knowing When to Look for What and Where: Evaluating Generation of Spatial Descriptions with Adaptive Attention

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, Box 200, 405 30 Gothenburg, Sweden
{mehdi.ghanimifard,simon.dobnik}@gu.se

Abstract. We examine and evaluate adaptive attention [17] (which balances the focus on visual features and focus on textual features) in generating image captions in end-to-end neural networks, in particular how adaptive attention is informative for generating spatial relations. We show that the model generates spatial relations more on the basis of textual rather than visual features and therefore confirm the previous observations that the learned visual features are missing information about geometric relations between objects.

Keywords: Image descriptions · Grounded neural language model Attention model · Spatial descriptions

1 Introduction

End-to-end neural networks are commonly used in image description tasks [17,28,29]. Typically, a pre-trained convolutional neural network is used as an encoder which produces visual features, and a neural language model is used as a decoder that generates descriptions of scenes. The underlying idea in this *representation learning* scenario [5] is that hidden features are learned from the observable data with minimum engineering effort of background knowledge. For example in word sequence generation only some general properties of a sequence structure [26] are given to the learner while the learner learns from the observed data what word to choose in a sequence together with a representation of features. Recent models such as [17,29] also add to the neural language model a model of visual attention over visual features which is inspired by the attention mechanism for alignment in neural machine translation [4]. It may be argued

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-11018-5_14) contains supplementary material, which is available to authorized users.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11132, pp. 153–161, 2019.
https://doi.org/10.1007/978-3-030-11018-5_14

that the attention mechanism introduces modularity to representation learning in the sense of *inception modules* [27] and *neural module networks* [2]. The visual attention is intended to detect the salient features of the image and align them with words predicted by the decoder. In particular, it creates a sum of the weighted final visual features at different regions of an image:

$$\boldsymbol{c}_t = \sum_{i=1}^{k} \boldsymbol{\alpha}_{ti} \boldsymbol{v}_i \tag{1}$$

where at time t, $\boldsymbol{c}_t$ represents the pooled visual features, i corresponds to k different regions of image, $\boldsymbol{v}_i$ is the visual representation of a particular region, and $\boldsymbol{\alpha}_{ti}$ represent the amount of attention on the specific region of the image. This representation provides the features for grounding the prediction of next word:

$$logPr(w_{t+1} = y_{t+1} | w_{1:t} = y_{1:t}, I = v_{1:k}) \approx f(y_{1:t}, c_t) \tag{2}$$

where f represents the end-to-end neural network for approximating the prediction of the next word in sentence.

However, not all words in natural language descriptions are directly grounded in visual features which leads [17] to extend the attention model [29] with an adaptive attention mechanism which learns to balance between the contribution of the visual signal and the language signal when generating a sequence of words.

$$\hat{\boldsymbol{c}}_t = \beta_t \boldsymbol{s}_t + (1 - \beta_t) \boldsymbol{c}_t \tag{3}$$

where at time t, $\hat{\boldsymbol{c}}_t$ is a combined representation of language features and visual features in addition to $\boldsymbol{c}_t$ of the visual features from Eq. 2. $\boldsymbol{s}_t$ is obtained from the memory state of the language model, and β_t ranging between $[0, 1]$ is the adaptive attention balancing the combination of vision and language features.

The performance of the image captioning systems when evaluated on the acceptability of the generated descriptions is impressive. However, in order to evaluate the success of learning we also need to understand better what the system has learned especially because good overall results may be due to the dataset artefacts or the system is simply learning from one modality only, ignoring the other [1]. Understanding the representations that have been learned also gives us an insight into building better systems for image captioning, especially since we do not have a clear understanding of the features in the domain. An example of work in this area is [15] which evaluates visual attention on objects localisation. [25] developed the FOIL dataset as a diagnostic tool to investigate if models look at images in caption generation. In [24] they examine the FOIL diagnostic for different parts-of-speech and conclude that the state of the art models can locate objects but their language models do not perform well on other parts-of-speech.

The current paper focuses on generation of spatial descriptions, in particular locative expressions such as "the chair to the left of the sofa" or "people close to the statue in the square". Spatial relations relate a target ("people") and landmark objects ("the statue") with a spatial relation ("close to"). They depend on several contextual sources of information such as scene geometry

("where" objects are in relation to each other), properties or function of objects and their interaction ("what" is related) as well as the interaction between conversational participants [8,10,11,13,21]. The features that are relevant in computational modelling of spatial language are difficult to determine simply by manually considering individual examples and they are normally identified through experimental work. The representation learning models are therefore particularly suited for their computational modelling.

However, the end-to-end vision and language models with attention are implemented in a way to recognise objects and localise their area in an image [3,18]. To generate spatial relations, [20] propose a combination of visual representations from convolutional neural networks and manually designed geometric representation of targets and landmarks. On quick examination, the representation of attention over images as in [29] gives an impression that attention captures both "what" and "where", especially because the attention graphs resemble *spatial templates* [16]. However, [12] argue that due to the design properties of image captioning networks, attention does not capture "where" as these models are built to identify objects but not geometric relations between them which they examine at the level of qualitative evaluation of attention on spatial relations.

In this paper we quantitatively evaluate the model of adaptive attention of [17] in predicting spatial relations in image descriptions. The resources used in our evaluation are described in Sect. 2. In Sect. 3 we examine the grounding of different parts-of-speech in visual and textual part of attention. Furthermore, in Sect. 4 we investigate the attention on spatial relations, targets and landmarks. We conclude by providing the possible directions for future studies and improvements.

2 Datasets and Pre-trained Models

As a part of their implementation [17] provide two different pre-trained image captioning models: Flickr30K [30] and MS-COCO [14].[1] We base our experiments on spatial descriptions of 40,736 images in the MS-COCO test corpus.

3 Visual Attention and Word Categories

Hypothesis. Our hypothesis is that visual attention in the end-to-end image captioning systems works as an object detector similar to [3,18]. Therefore, we expect the adaptive attention to prefer to attend to visual features rather than the language model features when predicting categories of words found in noun phrases that refer to objects, in particular head nouns. We expect that both scores will be reversed: more predictable words by the language model in the blind test receive less visual attention.

[1] https://filebox.ece.vt.edu/~jiasenlu/codeRelease/AdaptiveAttention.

Method. We use the pre-trained model of adaptive attention[2] to generate a description for each of the 40,736 images in the MS-COCO-2014 test. All the attention values are logged (α, β). We apply universal part-of-speech tagger from NLTK [6] on the generated sentences and report the average visual attentions on each part-of-speech. We match our results with results on the degree of predictability of each part-of-speech from the language model without looking at the image from the blind test of [24]. Note that we do not investigate the overall quality of the model on the test set (this has already been evaluated by its authors) but what kind of attention this model gives to vision and language features used to generate a word of each category. The evaluation code: https://github.com/GU-CLASP/eccv18-sivl-attention.

Results. Table 1 indicates that the highest degree of visual attentions is on numbers (NUM), nouns (NOUN), adjectives (ADJ) and determiners (DET) respectively. Pronouns (PRON) and particles (PRT) receive the lowest degree of visual attention. Verbs (VERB) and adverbs (ADV) are placed in the middle of this sorted list. Spatial relations which are mainly annotated as prepositions/adpositions (ADP) receive the second lowest visual attention, higher only than pronouns (PRON) and particles (PRT). Our results are different from the accuracy scores of detecting mismatch descriptions in the FOIL classification task [24]. For example, the model assigns predicts the mismatch on ADJ easier than mismatch on ADV. As hypothesised, the part-of-speech that make up noun phrases receive the highest visual attention (and the lowest language model attention). The results also indicate that the text is never generated by a single attention alone but a combination of visual and language model attentions. Since some spatial relations are often annotated as adjectives (e.g. "front"), a more detailed comparison on spatial terms is required.

4 Visual Attention When Grounding Spatial Descriptions

In generation of a sequence of words that make up a spatial description, which type of features or evidence is taken into consideration by the model as the description unfolds?

Hypothesis. In Sect. 3, we argued that the generation of spatial relations (prepositions/adpositions) is less dependent on visual features compared to noun phrases due to the fact that the learned visual features are used for object recognition and not recognition of geometric spatial relations between objects. Moreover, the visual clues that would predict the choice of spatial relation are not in one specific region of an image; this is dependent on the location of the target, the landmark and the configuration of the environment as a whole. Therefore, our hypothesis is that when generating spatial relations the visual attention is more spread over possible regions rather than being focused on a specific object.

[2] https://filebox.ece.vt.edu/~jiasenlu/codeRelease/AdaptiveAttention/model/COCO/coco_challenge/model_id1_34.t7.

Table 1. The average visual attention $(1 - \beta)$ for predicting words on each part-of-speech. The scores from the blind test indicate the accuracy of detecting a mismatch description in the FOIL-classification task [24].

POS	Count	Mean ± std	Blind test
NUM	1882	0.81 ± 0.08	-
NOUN	134332	0.78 ± 0.12	0.23
ADJ	23670	0.77 ± 0.14	0.76
DET	96641	0.73 ± 0.12	-
VERB	38381	0.70 ± 0.11	0.57
CONJ	6755	0.70 ± 0.13	-
ADV	184	0.69 ± 0.12	0.18
ADP	64332	0.62 ± 0.15	0.54
PRON	2347	0.53 ± 0.14	-
PRT	6462	0.52 ± 0.21	-

Method. The corpus tagged with POS from the previous section was used. In order to examine the attention on spatial relations, a list of keywords from [11,13] was used to identify them, provided that they have a sufficient frequency in the corpus. The average adaptive visual attention for each word can be compared with the scores in Table 1 for different parts-of-speech. In each sentence, the nouns before the spatial relation and the nouns after the spatial relations are taken as the most likely targets and landmarks respectively. The average adaptive visual attention on targets, landmarks and spatial relations is recorded.

Results. In Table 2 we report for each spatial relation and its targets and landmarks the average adaptive visual attention. The adaptive attentions for triplets are comparable with the figures for each part-of-speech in Table 1. In the current table, the variance of visual attentions is reported with the $max - min$ measure which is the difference between maximum and minimum attentions on a 7×7 plane representing the visual regions in the model. Lower values indicate either a low attention or a wider spread of attended area, hence less visual focus. Higher values indicate that there is more visual focus. For each spatial relation, the triplets must be compared with each other. In all cases, our hypothesis is confirmed: (1) the adaptive visual attention is lower on predicting spatial relations which means that they receive overall less visual attention, (2) with the exception of "under", the difference between maximum and minimum visual attentions are lower with spatial relations which means that the attention is spread more over the 7×7 plane. Figure 1 shows a visualisation of these results for "under" and "over". The results also show that landmarks in most cases receive less visual attention in comparison to targets. This indicates that after providing a target and a spatial relation, the landmark is more predictable from the language model (for a similar observation see [9]).

Table 2. The average score of adaptive visual attention for target (TRG) relation (REL) landmark (LND) triplets per each relation in the first column and the average difference between the highest and the lowest value of visual attention for the same items in the second column.

Descriptions Spatial relations	Average $(1-\beta_t)$ TRG, REL, LND	Average $(max(\hat{\alpha_t}) - min(\hat{\alpha_t}))$ TRG, REL, LND
Under	0.84, **0.73**, 0.79	0.0252, 0.0151, **0.0139**
Front	0.83, **0.70**, 0.82	0.0230, **0.0136**, 0.0154
Next	0.82, **0.68**, 0.78	0.0224, **0.0136**, 0.0138
Back	0.85, **0.68**, 0.84	0.0332, **0.0186**, 0.0272
In	0.82, **0.68**, 0.77	0.0250, **0.0149**, 0.0164
On	0.81, **0.68**, 0.75	0.0249, **0.0154**, 0.0175
Near	0.80, **0.67**, 0.76	0.0221, **0.0133**, 0.0169
Over	0.77, **0.62**, 0.75	0.0205, **0.0133**, 0.0193
Above	0.73, **0.64**, 0.77	0.0167, **0.0134**, 0.0231

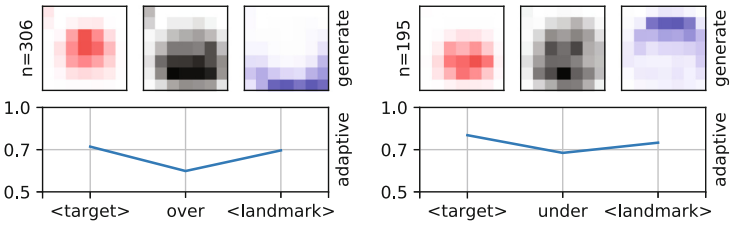

Fig. 1. Each square in a box in the first row represents an averaged attention for a location in the 7×7 grid over all n generated samples ($\hat{\alpha}$). The colours fade to white with lower values. The bottom graphs show their average over the entire plane, indicating the degree of adaptive visual attention $(1-\beta)$, also reported in Table 2.

5 Discussion and Conclusion

In this paper we explored to what degree adaptive attention is grounding spatial relations. We have shown that adaptive visual attention is more important for grounding objects but less important for grounding spatial relations which are not directly represented with visual features. As a result the visual attention is diffused over a larger space. The cause for a wider attended area can be due to high degree of noise in visual features or lack of evidence for visual grounding.

This is a clear shortcoming of the image captioning model, as it is not able to discriminate spatial relations on the basis of geometric relations between the objects, for example between relations such as "left" and "right". The future work on generating image descriptions therefore requires models where visual geometry between objects is explicitly represented as in [7]. The study

also shows that when generating spatial relations, a significant part of the information is predicted by the language model. This is not necessarily a disadvantage. The success of distributional semantics shows that language models with word embeddings can learn a surprising amount of semantic information without access to visual grounding. As mentioned in the introduction, spatial relations do not depend only on geometric arrangement of objects but also functional properties of objects. For example, [9] demonstrate that neural language models encode such functional information about objects when predicting spatial relations. Since, each spatial relation has different degree of functional and geometric bias [8], the adaptive attention considering visual features and textual features is also reflective of this aspect.

Models for explaining language model predictions such as [19] are also related to this study and its future work.

Our study focused on the adaptive attention in [17] which explicitly models attention as a focus on visual and language features. However, further investigations of other types of models of attention could be made and this will be the focus of our future work. We expect that different models of attention will behave similarly in terms of attending visual features on spatial relations because the way visual features are represented: they favour detection of objects and not their relative geometric arrangement. Our future work we will therefore focus on how to formulate a model to be able to learn such geometric information in an end-to-end fashion. Methodologies such as [22] and [23] which investigate the degree of effectiveness of features without attention are also possible directions of the future studies.

Acknowledgements. We are also grateful to the anonymous reviewers for their helpful comments on our earlier draft. 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

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: overcoming priors for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4971–4980 (2018)
2. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 39–48 (2016)
3. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. arXiv preprint arXiv:1412.7755 (2014)
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
5. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)

6. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media Inc., Sebastopol (2009)
7. Coventry, K.R., et al.: Spatial prepositions and vague quantifiers: implementing the functional geometric framework. In: Freksa, C., Knauff, M., Krieg-Brückner, B., Nebel, B., Barkowsky, T. (eds.) Spatial Cognition 2004. LNCS (LNAI), vol. 3343, pp. 98–110. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-32255-9_6
8. Coventry, K.R., Garrod, S.C.: Saying, Seeing, and Acting: The Psychological Semantics of Spatial Prepositions. Psychology Press, Hove (2004)
9. Dobnik, S., Ghanimifard, M., Kelleher, J.D.: 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, pp. 1–11. Association for Computational Linguistics, New Orleans, 6 June 2018
10. Dobnik, S., Kelleher, J.D.: Modular networks: an approach to the top-down versus bottom-up dilemma in natural language processing. In: Forthcoming in Post-proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML), vol. 1, no. 1, pp. 1–8, 12–14 June 2017
11. Herskovits, A.: Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English. Cambridge University Press, Cambridge (1986)
12. Kelleher, J.D., Dobnik, S.: What is not where: the challenge of integrating spatial representations into deep learning architectures. CLASP Papers in Computational Linguistics, p. 41 (2017)
13. Landau, B., Jackendoff, R.: "what" and "where" in spatial language and spatial cognition. Behav. Brain Sci. **16**(2), 217–238, 255–265 (1993)
14. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
15. Liu, C., Mao, J., Sha, F., Yuille, A.L.: Attention correctness in neural image captioning. In: AAAI, pp. 4176–4182 (2017)
16. Logan, G.D., Sadler, D.D.: A computational analysis of the apprehension of spatial relations. In: Bloom, P., Peterson, M.A., Nadel, L., Garrett, M.F. (eds.) Language and Space, pp. 493–530. MIT Press, Cambridge (1996)
17. Lu, J., Xiong, C., Parikh, D., Socher, R.: 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), vol. 6 (2017)
18. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems, pp. 2204–2212 (2014)
19. Park, D.H., Hendricks, L.A., Akata, Z., Schiele, B., Darrell, T., Rohrbach, M.: Attentive explanations: justifying decisions and pointing to the evidence. arXiv preprint arXiv:1612.04757 (2016)
20. Ramisa, A., Wang, J., Lu, Y., Dellandrea, E., Moreno-Noguer, F., Gaizauskas, R.: 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, pp. 214–220 (2015)
21. Regier, T.: The Human Semantic Potential: Spatial Language and Constrained Connectionism. MIT Press, Cambridge (1996)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)

23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626 (2017)
24. Shekhar, R., Pezzelle, S., Herbelot, A., Nabi, M., Sangineto, E., Bernardi, R.: Vision and language integration: moving beyond objects. In: IWCS 2017–12th International Conference on Computational Semantics–Short papers (2017)
25. Shekhar, R., et al.: FOIL it! find one mismatch between image and language caption. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Long Papers), vol. 1, pp. 255–265 (2017)
26. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
27. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
28. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156–3164. IEEE (2015)
29. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
30. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguist. **2**, 67–78 (2014)