Towards Natural Language Interfaces for Interacting with Remote Sensing Data*

Bruno Martins[1,2,3][0000−0002−3856−2936] and João Daniel Silva[1,3][0000−0001−6474−7822]

[1] INESC-ID, Lisbon, Portugal
[2] LUMLIS (Lisbon ELLIS Unit), Lisbon, Portugal
[3] Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal
{bruno.g.martins,joao.daniel.silva}@tecnico.ulisboa.pt

Abstract. Image captioning and visual question answering are exciting problems that combine natural language processing and computer vision, currently attracting a significant interest. Some previous efforts have looked into these problems in the context of remote sensing imagery, opening a wide range of possibilities in terms of human interaction with these data through natural language. Still, the components that are involved in previously proposed models can be significantly improved, and evaluation has also mostly been carried out on relatively small datasets, often built automatically and without much diversity. This vision paper briefly surveys the current state-of-the-art in vision and language methods dealing with remote sensing data, also discussing some of the open challenges and possibilities for future work.

Keywords: Remote Sensing Visual Question Answering · Remote Sensing Image Captioning · Learning with Vision and Language Inputs

DOI: https://doi.org/10.25436/E2S88R

1 Introduction

Image captioning and Visual Question Answering (VQA) are exciting problems that combine language processing and computer vision, currently attracting a significant interest. Some previous efforts have looked into these problems in the context of remote sensing imagery, arguing that these tasks can provide a useful framework to extract generic information from Earth observation data. One formulation for these problems involves taking as input an aerial image accompanied by a question (e.g., asking for a scene description, or asking about the presence of particular features or relational attributes), and producing as output a relevant answer (or description) based on the aerial image. Considering this formulation, one can imagine a language-based interface to a system like Google Earth, e.g. allowing different types of users to ask questions about the presence and quantity of particular objects within an

* This work was partially supported through Fundação para a Ciência e Tecnologia (FCT), namely through the FCT project grant with reference PTDC/CCI-CIF/32607/2017 (MIMU), as well as through the INESC-ID multi-annual funding from the PIDDAC programme with reference UIDB/50021/2020.

aerial scene (e.g., addressing questions like *"how many residential buildings are there in the image?" or "are there water bodies next to agricultural areas in the image?"*).

Previous studies have adapted techniques originally developed for general purpose images (i.e., techniques developed and tested over collections of ground-level photos), based on machine learning with deep neural networks. Authors have considered tasks such as remote sensing image captioning [29, 40, 10], VQA [40, 5], or even remote sensing image generation from textual prompts [37]. Although these previous studies have opened a wide range of possibilities, the model components that are involved can still be significantly improved (e.g., few previous studies have reported on approaches based on Transformer neural networks, which are currently the state-of-the-art for other vision and language tasks). Evaluation has also mostly been carried out on relatively small datasets, often built automatically (e.g., most datasets for remote sensing VQA were built from OpenStreetMap data, through rules and patterns for the automatic creation of image/question/answer triples involving vector objects and their properties) and without much diversity. To become truly general purpose, models should be trained with large amounts of data from different geographical areas, at the same time also aiming for different thematic objectives.

This paper briefly surveys the current state-of-the-art in terms of models supporting natural language interfaces for interacting with remote sensing data, also discussing some of the open challenges and possibilities for future work.

2 Current Approaches for Remote Sensing Image Captioning and Visual Question Answering

Studies on remote sensing image captioning have been heavily inspired by work on natural (i.e., ground-level) image captioning, following template-based methods that generate the captions through rules and patterns that combine object classes detected within the images [33], retrieval-based methods that use captions associated to similar images [14, 34], or encoder-decoder approaches based on deep neural networks [27, 23, 42, 16, 29, 43, 40, 46, 10]. These latter methods are currently the most common, and many previous studies follow an architecture combining convolutional, recurrent, and attention components. An encoder corresponding to a convolutional neural network (e.g, a ResNet model pre-trained on ImageNet) generates a representation for the image, and a decoder based on a recurrent neural network generates the caption word-by-word, at each step using neural attention to weight parts of the visual input according to relevance for the current prediction (i.e., the caption is generated auto-regressively, with the model repeatedly predicting each token conditioned on the previous tokens and on image characteristics weighted through neural attention). Most previous work has focused on the proposal of improved attention mechanisms, accounting with characteristics or remote sensing data such as the need to deal with visual features at different scales. The dataset that is currently most used to support evaluation experiments is RSICD [23], featuring 10921 images and five sentence descriptions per image. Some studies have nonetheless pointed to problems with these data, with Li et al. [16] releasing an updated version featuring revisions and typographical/grammatical corrections, and with Ramos and Martins [29]

pointing to further problems in terms of diversity and correct English writing. A very recent study has proposed the larger NWPU-Captions dataset [10], featuring 157,500 sentences produced by 7 experienced annotators, in association to 31,500 images with a wider coverage of complex scenes (although only a single relatively simple model has, to this date, been tested with this new dataset).

Some previous efforts have also looked into Visual Question Answering (VQA) in the context of remote sensing imagery [22, 21, 20, 44, 8, 19, 40, 5]. State-of-the-art approaches are also based on deep neural networks, with systems often following an architecture that features: an encoder component corresponding to a pre-trained convolutional neural network; another encoder corresponding to a recurrent neural network (e.g., an LSTM model combined with pre-trained word embeddings) that is used to generate a representation for the question; a fusion and classification component that combines both types of information and selects an answer from a set of candidates, thus treating the VQA problem as a multi-label classification task (i.e., the typical models are tasked with selecting the correct answer from a closed predefined set of possible answers). Existing evaluation datasets[4] were created through automated procedures, leveraging OpenStreetMap data or land coverage information to automatically derive the image/question/answer triplets. The questions and answers can refer to object counting, relative position reasoning, or scene classification, although the existing datasets again do not have much diversity.

3 Towards Larger Models Trained with More Data

In vision and language tasks involving remote sensing imagery, apart from a few very recent studies [9, 12, 30, 18, 5, 46], researchers have not explored recent architectures that replace the convolutional and recurrent encoders with modules leveraging self-attention operations (i.e., models using multi-modal Transformers). Previous work on other domains has showed that using Transformers pre-trained on a massive amount of image-caption pairs (e.g., models such as CLIP [28] or CoCa [38]) can bring significant advantages to both image captioning and VQA [32, 11, 4, 3]. One very recent study [5] has in fact leveraged CLIP for remote sensing VQA, combining the bi-modal features extracted with the CLIP model through a co-attention mechanism. Still, it is likely the case that a better adaptation of these methods, from general ground-level photos to the remote sensing domain, would require large datasets for better adjusting the image and text encoders. One previous project[5] has focused on adapting CLIP to aerial images, using captioning datasets such as RSICD [23] to adjust the model parameters through a contrastive learning objective that maximizes the similarity between representations for the images and the ground-truth corresponding captions. Still, experiments with the resulting model focused on zero-shot image classification (i.e., without any additional supervised training, the model is used to select a class according to the similarity between the class name and the input image). Perhaps the CLIP-RSICD model can also be further improved by considering other recent contrastive learning objectives besides the original one

[4] https://rsvqa.sylvainlobry.com
[5] https://github.com/arampacha/CLIP-rsicd

from CLIP, such as that from SimVLM [35], together with larger datasets (e.g., built automatically from resources such as OpenStreetMap, through templates for generating textual descriptions from the vector objects and their properties, or built by paraphrasing/transforming instances within existing datasets).

One interesting limitation of state-of-the-art visual encoders relates to the relatively low resolution of the input images, leading to a loss of fine-grained spatial detail. Although the existing remote sensing VQA datasets were constructed in ways that explored the resolution of the images (i.e., low resolution datasets feature large scale queries related to distinctions such as rural/urban, while high resolution datasets include questions relating to small objects like houses or trees), most previous studies have shown that good results can still be achieved with only low-resolution inputs (i.e., the original images are often reduced to inputs featuring 224×224 pixels). Future work can perhaps consider the use of higher resolution inputs, checking if (and in which cases) the results can be improved. One can for instance consider concatenating the representations produced for different patches from the original high-resolution images, or alternatively consider vision Transformers featuring multi-scale structures [39, 41] as the image encoders.

4 Towards Models that are Right for the Right Reasons

When addressing remote sensing image captioning or VQA, we expect models to generate truthful predictions, based on the evidence within the image contents and based on the actual question intention. Unfortunately, instead of *sticking to the facts*, current approaches can rely on spurious correlations, and often follow biases induced by the training data and/or the models. Previous studies in remote sensing VQA have for instance shown that models are brittle to linguistic variations in questions/answers, or instead that models can often infer the correct answers just by analyzing common patterns in the questions.

Some previous studies focusing on general ground-level photos have proposed approaches to deal with the aforementioned limitations, based on cyclic consistency, contrastive learning, counterfactual inference, or data augmentation, to improve model robustness [2, 24, 17]. One can for instance automatically generate new entailed questions, that should produce a consistent answer, or counterfactual questions that should lead to a different answer. Similarly, one can edit the input images semantically, to create different variants where the same question-answer pair holds (e.g., removing irrelevant objects in a way such that the answer remains invariant). Still, these techniques have yet to be applied to remote sensing imagery. We believe there is significant potential in these approaches for improving current models, given that resources like OpenStreetMap can be effectively used to build challenging training examples featuring either compatible or counterfactual properties.

Another possible strategy to circumvent the reliance on spurious correlations, which can also lead to improvements in terms of model interpretability, relates to improving the localization capabilities of remote sensing vision and language models (i.e., more than producing the correct answer, these models should also correctly point to the image regions that explain the answer). Using more powerful visual encoders can perhaps contribute in this direction, given that previous studies using self-supervised contrastive learning have shown that the features resulting from

vision Transformers contain explicit information about semantic object boundaries within the input images [6, 45]. This behavior can be further reinforced, through multi-task learning methods that simultaneously optimize the results for VQA or image captioning, and the generation of segmentation masks that explain the generated output [15, 13, 31, 43] (e.g., for a question focusing on the number of buildings present in a given image, one can train the VQA model to simultaneously generate a segmentation mask focusing on the buildings, while attending to these same regions for producing the answer). Again, since OpenStreetMap can be used for the creation of training datasets, it should be fairly straightforward to complement image/question/answer triples with the appropriate segmentation maps derived from the vector data, to guide model training.

5 Towards Open Domain and Conversational Models

Treating remote sensing VQA as a classification problem is perhaps too artificial, severely limiting the answers that are produced, and consequently also the questions that can be addressed. In fact, current models have only been trained and tested with datasets involving a small number of answer types (e.g., for questions relating to counting features, the answers are typically binned into a small set of classes).

We argue that research in the area should ideally move from closed into open settings: VQA models should generate the text of the answers word-by-word, as in image captioning tasks and without restricting the type of answer that can be generated, rather than select the answer class from a closed pre-defined set. Towards this goal, we see significant potential in joining the datasets and methods for remote sensing VQA and image captioning, bringing together these two problems [36]. The existing datasets can easily be combined to support the training of open-ended models (e.g., image captioning data can be extended with elicitation prompts such as *"what is being shown in this image"*, and combined with VQA data extended to consider different lexical realizations for the answers), and new datasets can also be automatically created, again from resources such as OpenStreetMap, or by augmenting and adapting other existing datasets (e.g., questions can be derived from existing captions [7], and new captions can be constructed from templates similar to those used for creating questions in the existing remote sensing VQA datasets).

One should ideally also consider conversational scenarios, in which questions or prompts for descriptions are given in the context of a dialog history. This would correspond to more natural user interactions [25, 26, 1, 3], although addressing the task requires different strategies for the construction of training and evaluation datasets, featuring multi-turn conversations about the contents of remote sensing imagery.

6 Conclusions

We briefly surveyed vision and language methods dealing with remote sensing data, and we discussed open challenges and possibilities for future work, specifically focusing on (i) the training of larger models with more data and larger input images, (ii) improving model robustness, and (iii) moving from closed into open settings, in which the outputs are less restricted. We hope this discussion can inspire researchers to tackle the hard challenges in the area, aiming at real and significant progress.

References

1. Agarwal, S., Bui, T., Lee, J.Y., Konstas, I., Rieser, V.: History for visual dialog: Do we really need it? arXiv:2005.07493 (2020)
2. Agarwal, V., Shetty, R., Fritz, M.: Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. arXiv:2204.14198 (2022)
4. Barraco, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: The unreasonable effectiveness of CLIP features for image captioning: An experimental analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (2022)
5. Bazi, Y., Al Rahhal, M.M., Mekhalfi, M.L., Zuair, M., Melgani, F.: Bi-modal transformer-based approach for visual question answering in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing (2022)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE International Conference on Computer Vision (2021)
7. Changpinyo, S., Kukliansky, D., Szpektor, I., Chen, X., Ding, N., Soricut, R.: All you may need for VQA are image captions. arXiv:2205.01883 (2022)
8. Chappuis, C., Lobry, S., Kellenberger, B., Saux, B.L., Tuia, D.: How to find a good image-text embedding for remote sensing VQA? arXiv:2109.11848 (2021)
9. Chappuis, C., Zermatten, V., Lobry, S., Le Saux, B., Tuia, D.: Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering. In: Workshops of the IEEE Conference on Computer Vision and Pattern Recognition (2022)
10. Cheng, Q., Huang, H., Xu, Y., Zhou, Y., Li, H., Wang, Z.: NWPU-Captions dataset and MLCA-Net for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing (2022)
11. Eslami, S., de Melo, G., Meinel, C.: Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? arXiv:2112.13906 (2021)
12. Gajbhiye, G.O., Nandedkar, A.V.: Generating the captions for remote sensing images: A spatial-channel attention based memory-guided transformer approach. Engineering Applications of Artificial Intelligence **114** (2022)
13. Gan, C., Li, Y., Li, H., Sun, C., Gong, B.: VQS: Linking segmentations to questions and answers for supervised attention in VQA and question-focused semantic segmentation. In: IEEE International Conference on Computer Vision (2017)
14. Hoxha, G., Melgani, F., Slaghenauffi, J.: A new CNN-RNN framework for remote sensing image captioning. In: IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (2020)
15. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv:2201.03546 (2022)
16. Li, Y., Fang, S., Jiao, L., Liu, R., Shang, R.: A multi-level attention model for remote sensing image captions. Remote Sensing **12**(6) (2020)
17. Liang, Z., Jiang, W., Hu, H., Zhu, J.: Learning to contrast the counterfactual samples for robust visual question answering. In: ACL Conference on Empirical Methods in Natural Language Processing (2020)
18. Liu, C., Zhao, R., Shi, Z.: Remote-sensing image captioning based on multilayer aggregated transformer. IEEE Geoscience and Remote Sensing Letters **19** (2022)

19. Lobry, S., Demir, B., Tuia, D.: RSVQA Meets BigEarthNet: A New, Large-Scale, Visual Question Answering Dataset for Remote Sensing. In: IEEE International Geoscience and Remote Sensing Symposium (2021)
20. Lobry, S., Marcos, D., Kellenberger, B., Tuia, D.: Better Generic Objects Counting When Asking Questions to Images A Multitask Approach for Remote Sensing Visual Question Answering. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2020)
21. Lobry, S., Marcos, D., Murray, J., Tuia, D.: RSVQA: Visual Question Answering for Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing **58** (2020)
22. Lobry, S., Murray, J., Marcos, D., Tuia, D.: VQA From Remote Sensing Images. In: IEEE International Geoscience and Remote Sensing Symposium (2019)
23. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring Models and Data for Remote Sensing Image Caption Generation. IEEE Transactions on Geoscience and Remote Sensing **56** (2018)
24. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual VQA: A cause-effect look at language bias. In: IEEE Conference on Computer Vision and Pattern Recognition (2021)
25. Niu, Y., Zhang, H., Zhang, M., Zhang, J., Lu, Z., Wen, J.R.: Recursive visual attention in visual dialog. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
26. Qi, J., Niu, Y., Huang, J., Zhang, H.: Two causal principles for improving visual dialog. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
27. Qu, B., Li, X., Tao, D., Lu, X.: Deep semantic understanding of high resolution remote sensing image. In: International Conference on Computer, Information and Telecommunication Systems (2016)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 (2021)
29. Ramos, R., Martins, B.: Using neural encoder-decoder models with continuous outputs for remote sensing image captioning. IEEE Access **10** (2021)
30. Ren, Z., Gou, S., Guo, Z., Mao, S., Li, R.: A mask-guided transformer network with topic token for remote sensing image captioning. Remote Sensing **14**(12) (2022)
31. Sharma, V., Bishnu, A., Patel, L.: Segmentation guided attention networks for visual question answering. In: ACL Student Research Workshop (2017)
32. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can CLIP benefit vision-and-language tasks? arXiv:2107.06383 (2021)
33. Shi, Z., Zou, Z.: Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? IEEE Transactions on Geoscience and Remote Sensing **55** (2017)
34. Wang, B., Zheng, X., Qu, B., Lu, X.: Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13** (2020)
35. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: SimVLM: Simple visual language model pretraining with weak supervision. arXiv:2108.10904 (2021)
36. Wu, J., Hu, Z., Mooney, R.J.: Generating question relevant captions to aid visual question answering. arXiv:1906.00513 (2019)
37. Xu, Y., Yu, W., Ghamisi, P., Kopp, M., Hochreiter, S.: Txt2Img-MHN: Remote sensing image generation from text using modern Hopfield networks. arXiv:2208.04441 (2022)
38. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive captioners are image-text foundation models. arXiv:2205.01917 (2022)

39. Yu, Q., Xia, Y., Bai, Y., Lu, Y., Yuille, A.L., Shen, W.: Glance-and-gaze vision transformer. In: Annual Meeting on Neural Information Processing Systems (2021)
40. Yuan, Z., Mou, L., Wang, Q., Zhu, X.X.: From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing **60** (2022)
41. Zhang, P., Dai, X., Yang, J., Xiao, B., Yuan, L., Zhang, L., Gao, J.: Multi-scale vision Longformer: A new vision transformer for high-resolution image encoding. In: IEEE Conference on Computer Vision and Pattern Recognition (2021)
42. Zhang, X., Wang, X., Tang, X., Zhou, H., Li, C.: Description generation for remote sensing images using attribute attention mechanism. Remote Sensing **11**(6) (2019)
43. Zhao, R., Shi, Z., Zou, Z.: High-resolution remote sensing image captioning based on structured attention. IEEE Transactions on Geoscience and Remote Sensing **60** (2021)
44. Zheng, X., Wang, B., Du, X., Lu, X.: Mutual Attention Inception Network for Remote Sensing VQA. IEEE Transactions on Geoscience and Remote Sensing **60** (2022)
45. Zhou, C., Loy, C.C., Dai, B.: DenseCLIP: Extract free dense labels from clip. arXiv:2112.01071 (2021)
46. Zia, U., Riaz, M.M., Ghafoor, A.: Transforming remote sensing images to textual descriptions. International Journal of Applied Earth Observation and Geoinformation **108** (2022)