RSVQA MEETS BIGEARTHNET: A NEW, LARGE-SCALE, VISUAL QUESTION ANSWERING DATASET FOR REMOTE SENSING

Sylvain Lobry[1]*, Begüm Demir*[2]*, Devis Tuia*[3]

[1] LIPADE, Université de Paris, Paris, France
[2] Technische Universität Berlin, Berlin, Germany
[3] Ecole Polytechnique Fédérale de Lausanne, Sion, Switzerland

ABSTRACT

Visual Question Answering is a new task that can facilitate the extraction of information from images through textual queries: it aims at answering an open-ended question formulated in natural language about a given image. In this work, we introduce a new dataset to tackle the task of visual question answering on remote sensing images: this large-scale, open access dataset extracts image/question/answer triplets from the BigEarthNet dataset. This new dataset contains close to 15 millions samples and is openly available. We present the dataset construction procedure, its characteristics and first results using a deep-learning based methodology. These first results show that the task of visual question answering is challenging and opens new interesting research avenues at the interface of remote sensing and natural language processing. The dataset and the code to create and process it are open and freely available on `https://rsvqa.sylvainlobry.com/`

Index Terms— Visual Question Answering, Dataset, Deep learning, Natural Language, Computer Vision

1. INTRODUCTION

Visual Question Answering (VQA) is a task introduced in the computer vision community [1] that aims at answering an open-ended question relative to an image, formulated in natural language. In a remote sensing context, this task can be used as a framework to extract generic information from remote sensing images, as proposed by [2]. By using natural language as an interface, technical skills in computer vision would no longer be necessary to extract specific information from remote sensing images.

Natural language can be used in conjunction with remote sensing images for instance to generate captions (or descriptions). To this effect [3] encodes the image to a latent space and decodes it to generate a caption. [4] aims at solving the problem of redundant information in ground truth caption by summarizing them. In addition to VQA, natural language can also be used to interact with remote sensing images through image querying: this has been approached through fixed rules by [5] allowing to query images based on their meta-data or by generating a description of an image which is compared to the query in a latent space in [6].

VQA, on the other hand, allows to extract high-level information from the image content. VQA models are generally based on deep learning [7] and therefore call for large datasets. Two datasets have been proposed for VQA for remote sensing (RSVQA), in [2]. The first one is based on nine Sentinel-2 tiles covering the Netherlands, and the second on $10'659$ high-resolution (15cm) patches covering parts of the USA. Using information from OpenStreetMap, $77'232$ and $1'066'316$ image/question/answer triplets have been constructed, respectively. These two RSVQA datasets show two main limitations: the number of possible answers is limited (9 and 98, respectively) and the number of samples is not sufficient to train large deep learning architectures.

As deep learning models are being increasingly used to extract information from remote sensing data, there is a need for large-scale datasets. To this effect, [8] introduced the BigEarthNet dataset: it contains $590'326$ Sentinel-2 patches with labels extracted from the 2018 CORINE Land Cover (CLC) database. Each image from this dataset is annotated with the land cover classes from the 3rd level (L3, out of L1, L2 and L3) of the CLC hierarchy.

In this work, we create a new, large-scale, RSVQA dataset from the Sentinel-2 images and land cover classes of the BigEarthNet data named RSVQAxBEN. We describe the construction procedure and analyze the characteristics of this new dataset in section 2. Finally, we propose a baseline to tackle the VQA task and analyze the results obtained on our dataset in section 3. In support of reproducible research and future developments, the dataset and the code to create and process it are freely available on `https://rsvqa.sylvainlobry.com/`.

2. DATASET CREATION

The new dataset is composed of a series of image / question / answer triplets extracted from the Sentinel-2 images and land cover classes from BigEarthNet. In this section, we describe the construction of this dataset.

2.1. Question construction procedure

The procedure to construct a single question/answer pair for a given Sentinel-2 image and based on the CLC L3 labels provided by BigEarthNet is stochastic. Note that the only input to this procedure is the list of labels: the image is not used explicitly to construct the question/answer pairs.

We first create a list of present/absent labels at the three CLC levels and randomly choose a type of question to ask: it can either be a *yes/no question* about the presence of a land cover (e.g. "Is there a water body in the image?") or a question to which the answer is one or several land cover (LC) classes names (*Land cover questions*):

- *Yes/no questions*: first, the CLC level to which the question will apply is selected, followed by the answer (*yes* or *no*), directly retrieved by comparing the selected class with those present in the footprint of the image considered by a geographical query. Note that we select the number of land cover classes to be contained in the question (from one to three). When more than one element is selected, logical connectors (*and* and *or*) are used to build the question.
- *Land cover questions*: similarly to *yes/no questions*, the CLC level of the question is selected first. However, in this case, a question can also be defined on all three CLC levels. A question is then randomly chosen between:
  - a question asking which land cover classes are present;
  - a question, for which one or two land cover classes present in the image are given, and asking which other classes are present. In this case, the land cover classes provided in the question are chosen randomly from the list of classes present in the image.

To prevent the construction of ill-posed questions formulations, land cover classes containing the words "and" or "or" (e.g. "scrub and/or herbaceous vegetation associations") can not appear in the questions (as these words can be used as logical connectors between different classes). Using this procedure, we can define a large variety of questions, the answer of which can be automatically retrieved from the list of CLC labels associated to the image. Samples obtained with this procedure are shown in Table 1.

2.2. Dataset construction

Based on the question construction procedure, we generate a database for the $590'326$ patches of the BigEarthNet S2 dataset. We first select the RGB bands of each patch and rescale them to 8 bits (from 12). Then, we create 25 different questions for each patch, for a total of $14'758'150$ image/question/answer triplets.

Type	# classes	Question
yes/no	1	Are some **rice fields** present?
yes/no	2	Are there **inland wetlands** and **maritime wetlands**?
yes/no	3	Are there **water bodies** and **agricultural areas** or **wetlands** in the image?
LC	0	Which L1 classes are in the image?
LC	1	In addition to **water bodies**, which classes are in the image?
LC	2	Besides **transitional woodland/shrub** and **mixed forest**, what land cover classes are in the scene?

Table 1. Samples of questions from our dataset. Land cover classes in the question are in **bold**.

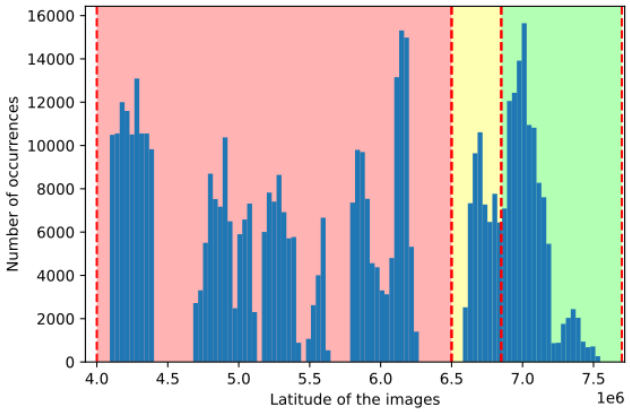

Fig. 1. Dataset splits relatively to the latitude of the images: the training set is indicated in red, the validation set in yellow and the test set in green.

2.3. Dataset splitting

When using supervised machine learning procedures, it is important to split the dataset in training, validation and test sets. We propose to split the dataset based on the spatial location of the image, to avoid positive biases due to geographical proximity of train and test samples. Specifically, we split the dataset according to their latitude, as shown in Figure 1. We also exclude from these splits the $70'987$ patches fully covered by seasonal snow, cloud or shadow following the recommendation of [8]. Using this methodology, we select 66% of the *valid* (i.e. not fully covered by snow, cloud or shadow) samples for training, 11% for validation and 23% for test.

2.4. Dataset analysis

Our dataset contains two types of questions: those for which the answer is "yes" or "no" (queries about the presence of one or several land covers) and those for which the answer is a (list

of) land cover type(s). *Yes/No questions* are the most frequent, with a frequency of 80.7%. This is because many of the land cover questions generated are ill-posed (see Section 2.1) and therefore discarded. Moreover, we observed $28'049$ possible answers among the $2'846'757$ *land cover questions* and this led to a a strong imbalance in the frequency of the single possible answers: the most frequent answer is *"None"*, covering 14.7% of the cases and the second one, *"Water bodies"*, is the answer to 3.6% of *land cover questions*. At the other end of the distribution, there are $12'218$ answers appearing only once in the dataset (an example of such an answer is *"Broad-leaved forest, complex cultivation patterns, coniferous forest, discontinuous urban fabric, land principally occupied by agriculture, with significant areas of natural vegetation, mixed forest, non-irrigated arable land and water courses"*).

Regarding the complexity of the questions, 72.3% contains at least one logical connector ("and" or "or"), with 27.1% of the questions containing two logical connectors.

3. BASELINE

3.1. Proposed model

To create a baseline for our dataset, we used the VQA model proposed in [2]. This model contains a feature extractor for the image (ResNet-152 [9] pre-trained on ImageNet [10]) and one for the question (skip-thoughts architecture [11], pre-trained on the BookCorpus dataset [12]). Each feature extractor produces a $1'200$ dimensional feature vector, and the two vectors are then merged with a point-wise multiplication and passed to a multi-layer perceptron for the prediction of the most probable answer (among a set of pre-defined ones). For a detailed description of our baseline, we refer the reader to [2].

3.2. Training settings

Output's dimension: Since the number of possible answers in our training set is large ($26'875$ unique answers), we restrict the set of possible answers to the $1'000$ most frequent ones. These answers cover 98.1% of the answer space in the training set. Therefore, the output size of the multi-layer perceptron is $1'000$. In other words, the maximum accuracy that our model can reach on the training set is 98.1%.
Training: We use the optimizer proposed by [13] with a learning rate of 10^{-6} and a batch size of $1'024$ for 10 epochs. These hyper-parameters have been set experimentally.

3.3. Results

We present the results obtained with our baseline in Table 2. The accuracy is defined as the ratio between the number of correct answers and the number of questions. It can be seen that, while the performance on *yes/no questions* is on-par with performances reported in [2], *land cover questions* show a poor performance. This can easily be explained by looking at the confusion matrix in Figure 2. It can be seen that the model actually restrict its answer space to 17 answers, including the most frequent ones. This partially comes from the strong imbalance of the answer frequency: these 17 most frequent answers cover 87% of the answer space.

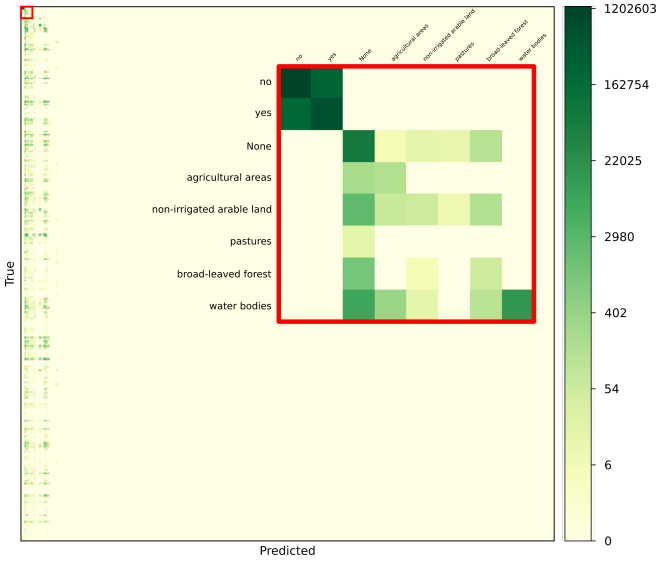

Fig. 2. Confusion matrix obtained with our proposed model in logarithm scale with a zoom (indicated in red) on the part covering the 8 most frequent answers.

Type of answer	Accuracy
Yes/No	79.92%
Land cover	20.57%
Global	69.83%

Table 2. Accuracy obtained with our proposed model.

Visually, we show in Figure 3 some predictions made by the baseline on the test set. In Figures 3(a, b), we see examples of good predictions over simple *yes/no questions*. Queries combining several land cover classes through logic connectors are also often answered correctly, as shown in Figure 3(c), where a water body is indeed present. However, we show in Figure 3(d) a question where our model does not answer correctly. This could be explained by the fact that logical formulas are not handled explicitly and the model only answers "yes" because mixed forests are present in the image. Finally, we show results on *Land cover questions* in Figures 3(e, f). In Figure 3(e), the model answer correctly to the question, where the answer is simple. However, Figure 3(f) shows that the model, despite being explicitly asked for L2 classes, only provides L1 classes. This can be explained when looking at the confusion matrix: the model can

Fig. 3. Visual results from the test set.

only predict simple combination of land cover classes, and this L2 combination is not frequent (it represents 0.005% of the answers in the test set). Improving the capacity of the model to answer complex combinations of classes is therefore a work that needs to be tackled in the future.

4. CONCLUSION

We introduced a new, large-scale dataset for RSVQA derived from the BigEarthNet data: RSVQAxBEN. In addition to a larger number of samples, this new dataset introduces new objects of interest (land cover classes) with a new form of complexity (logical formulas). We obtained encouraging results using a simple VQA architecture. However, we have shown that the model is not capable of dealing with the large imbalance of the dataset and complex, logical questions. These aspects should be explored in future works.

5. REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *ICCV*, 2015, pp. 2425–2433.

[2] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE TGRS*, vol. 58, no. 12, pp. 8555–8566, 2020.

[3] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE TGRS*, vol. 56, no. 4, pp. 2183–2195, 2017.

[4] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization driven deep remote sensing image captioning," *IEEE TGRS*, in press.

[5] X. Xuan, J. Liu, and J. Yang, "Research on the natural language querying for remote sensing databases," in *International Conference on Computer Science and Service System*, 2012, pp. 228–231.

[6] G. Hoxha, F. Melgani, and B. Demir, "Toward remote sensing image retrieval under a deep image captioning perspective," *IEEE JSTARS*, vol. 13, pp. 4462–4475, 2020.

[7] Q. Wu, D. Teney, P Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *CVIU*, vol. 163, pp. 21–40, 2017.

[8] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A Large-Scale Benchmark Archive For Remote Sensing Image Understanding," in *IGARSS*, 2019, pp. 5901–5904.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[11] R. Kiros, Y. Zhu, R. Salakhutdinov, R.S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," *NeurIPS*, 2015.

[12] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *ICCV*, 2015, pp. 19–27.

[13] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.