Multi-task prompt-RSVQA to explicitly count objects on aerial images

Christel Chappuis[1]
christel.chappuis@epfl.ch
Charlotte Sertic[2]
charlotte.sertic@epfl.ch
Nicola Santacroce[2]
nicola.santacroce@epfl.ch
Javiera Castillo Navarro[1]
javiera.castillonavarro@epfl.ch
Sylvain Lobry[3]
sylvain.lobry@u-paris.fr
Bertrand Le Saux[4]
bertrand.le.saux@esa.int
Devis Tuia[1]
devis.tuia@epfl.ch

[1] Environmental Computational Science and Earth Observation Laboratory
École Polytechnique Fédérale de Lausanne
Sion, Switzerland

[2] École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

[3] Laboratoire d'Informatique Paris Descartes (LIPADE)
Université Paris Cité
Paris, France

[4] Φ-lab
European Space Agency
Frascati, Italy

Abstract

Introduced to enable a wider use of Earth Observation images using natural language, Remote Sensing Visual Question Answering (RSVQA) remains a challenging task, in particular for questions related to counting. To address this specific challenge, we propose a modular Multi-task prompt-RSVQA model based on object detection and question answering modules. By creating a semantic bottleneck describing the image and providing a visual answer, our model allows users to assess the visual grounding of the answer and better interpret the prediction. A set of ablation studies are designed to consider the contributions of different modules and evaluation metrics are discussed for a finer-grained assessment. Experiments demonstrate competitive results against literature baselines and a zero-shot VQA model. In particular, our proposed model predicts answers for numerical *Counting* questions that are consistently closer in distance to the ground truth.

1 Introduction

Earth observation (EO) data has grown significantly over the past decades and, thanks to the sheer amount of imaging sensors available, has become an important source of knowledge for monitoring the planet. However, the usability of this information remains restrained by the technical requirements necessary for its extraction from the raw EO data [26]. Motivated by this challenge, the task of Remote Sensing Visual Question Answering (RSVQA [19])

was proposed. Using natural language, the goal of RSVQA is to predict an answer from a question about an EO image. While the technical principle is similar to that of Visual Question Answering (VQA) [2], motivations and end-users differ considerably. With RSVQA, the goal is to help scientists, decision makers or journalists, who often lack the technical skills required to run deep learning models, to extract valuable information from EO images. While accurate predictions are essential, users in RSVQA would benefit from being able to interpret and judge the answer in the context of the input image as well, and need to understand where the prediction originates from in terms of visual cues. We refer to this concept as *visual grounding*, i.e. ensuring that the answer is rooted in the image. An objective of this work is to orient RSVQA research towards more interpretable systems, where users have a direct insight on the information used by the model.

To reach this goal, we present a Multi-task prompt-RSVQA model that uses multiple visual descriptors, in particular object detection, to provide textual context for a BERT-like language model to answer EO-related questions. As the majority of published RSVQA methods are framed as a classification task, models have a tendency to learn the most probable or frequent answers based on the image and question features. We argue that a different strategy and evaluation is needed for numerical questions, in particular counting ones, which are prevalent and especially challenging in RSVQA datasets. Thus, in addition to the traditional classification head, a question answering module is used to process counting questions more reliably, i.e. numerically closer to the ground truths. A question type classifier decides which head is to be used. We use regression metrics in our evaluations, to keep track of the importance of errors, which is lost in classification settings (e.g. predicting 1 or 100 when the right answer is 2 has the same penalisation). Finally, we increase interpretability of results by producing visual response in addition to the textual one. Overall, our experiments show that the proposed model is competitive with "black-box" models, and able to predict more reliable answers to counting questions, while providing semantic bottlenecks to interpret the model predictions.

2 Related works

VQA. Earlier works in VQA achieved the highest performances with Transformer architectures and self-supervised strategies for large-scale pre-training, allowing the model to tackle several vision-and-language tasks (VQA, captioning [32], retrieval [5], commonsense knowledge [4], etc.). In terms of modular strategies, neural module networks [1] was pioneer, decomposing the answer prediction into a series of simple actions. Models for zero-shot predictions have been proposed such as Plug-and-Play VQA [25], where the answer prediction relies on three steps: matching question and image regions, captioning of the selected regions and a question answering module.

RSVQA. After its initial proposition [19], RSVQA focused on attention mechanisms [36], fusion strategies [6], and on tackling different levels of complexity via curriculum learning [33]. Visual transformers are increasingly considered as feature extractor [3] or as a fusion strategy [24]. The concept of describing images with keywords (context) and predicting answers using a language-only transformer model was investigated in Prompt-RSVQA [8]. However, Prompt-RSVQA was specifically tailored for land cover classification questions and did not perform well for other question types, in particular those about counting, which are challenging and frequent in RSVQA datasets [19, 21, 36]. Other works focused on the

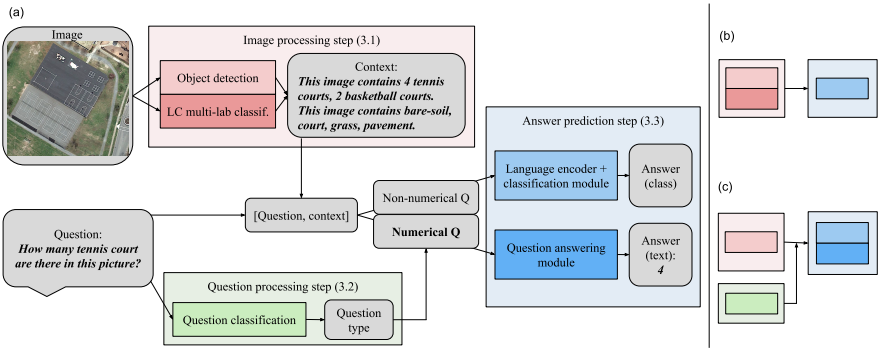

Figure 1: (a), Multitask prompt-RSVQA. The model is split into three steps (image processing, question processing and answer prediction) and five modules (objects detection, LC multi-label classification, question classification, language encoder + classification and question answering modules). (b-c) Ablation studies in this paper: (b) language encoder + classification module only in the answer prediction step (*Ablation no QA*) and (c) using object detection only in the image processing step (*Ablation no LC*)

issue of numerical questions: [18] proposed a multi-head strategy combining classification and regression, while [13] integrated an object detection model in the architecture. In this work, we build upon these recent advances [8, 13, 18] and propose a modular RSVQA system that first describes the image and then addresses different types of questions in a multi-task fashion. In particular, object detection and a question answering head are used to better process counting questions and grounding the answers in the visual modality.

3 Method

Our proposed "Multi-task prompt-RSVQA" first explicitly extracts the content of EO images using object detection and land cover classification and then transcribes their outputs into a context paragraph. Depending of the question type predicted by a question module, a specialised language model comes into play: one classifying question and context into answers (as in traditional RSVQA) and a second, a NLP question answering module, dealing with numerical questions by extracting the answer and highlighting it in the context (Figure 1(a)).

The model can be divided into three main steps, detailed in the next sections: (1) First the *image processing step*, where useful information (objects or land cover (LC) types), is extracted from the remote sensing image and transformed into a textual context. This first step consists of two modules: the objects detection module and the LC multi-label classification module, detailed in Section 3.1. (2) Secondly, *the question processing step*, where a Transformer language model classifies the type of question (Section 3.2). (3) Finally, *the answer prediction step*: depending on the question type, a classification module containing a language encoder, or a question answering module is activated (Section 3.3). While some of the different modules are fine-tuned individually, the full pipeline is run at inference.

3.1 Image processing step

Objects detection module. This module uses two object detectors with complementary categories: an oriented-RCNN [30] (pre-trained on the DOTA [12] and HRSC2016 [17] datasets) and a Faster-RCNN [22] pre-trained on COCO [15] and fine-tuned on the HRRSD dataset [35]. Object detectors output a list of detected objects, containing at least: the bounding boxes, the prediction of the object class and its probability score. To build our image context, we count in the list the number of occurrences detected above a defined probability threshold (0.3) for each type of object. Then, we create a text variable starting with "This image contains", and add one after the other the number of occurrences together with the object label, separated by commas (see Figure 1). Bounding boxes, not directly used in the image context, are kept in memory to produce the visual answer.

LC multi-label classification module. We describe LC, a key type of characterization for EO images, with a multi-label ResNet-18 [14] classifier trained on the multi-label version of the UCMerced dataset [9], which has 17 classes, such as *field*, *sand*, *sea*, *trees*, etc. As a multi-label model, none to several classes can be activated on each image. If at least one class is activated, we add a new text to the image context initially created by the object detection module, starting again with "This image contains ". Then, the label(s) of activated class(es) are appended one after the other with comma separation (see Figure 1). As we are dealing with "stuff" rather than "things" here, there is no number of occurrences.

3.2 Question processing step

To determine the type of question, we use a DistilBERT [23] encoder with fully-connected layers for Text Classification. DistilBERT is a distilled version of BERT [11], a widely used language model, with only half the number of attention layers (6 instead of 12). As a text classification task, this model is fine-tuned on RSIVQA to classify questions by their type: *Counting*, *Presence*, *Scene* and *Location*. To determine the question type, only the question text is used. The question is first tokenized (each element is transformed into a numerical dictionary value). The sequence is fed into DistilBERT: first through a word embedding step to project the numerical tokens into vector representations, then through the attention layers. The language encoder outputs a vector of 768 dimensions, that is then sent to a fully-connected layer to be projected onto a vector of dimension equal to the number of classes, i.e. the possible question types.

3.3 Answer prediction step

In this step, the question text is combined with the context extracted by the image processing step described in Section 3.1. Depending on the question type predicted by the question processing step, either of the modules below is used.

Answer classification module. When the question is considered a classification one (*presence, scene* and *location*), the {question, image context} pair of sequences is inputted to a Transformer language encoder. Question and context are tokenized together, and separated by a separation token. We again use DistilBERT and fine-tune it on our task, to output all possible non-numerical answer classes in the dataset.

Question answering (QA) module. For numerical questions (*counting*), the {question, image context} pair of sequences is fed through a question answering (QA) model, based on RoBERTa (Robustly Optimized BERT [16]) without any fine-tuning. RoBERTa is an extension of BERT which is trained on more data for a longer period of time. The QA model uses extractive question answering: it extracts the answer to the first sequence (question) from the second sequence (image context) as a text output along with its token position in the sequence. We rely on the QA model to identify the element the question refers to, and then focus on and extract only the number of occurrences associated with that element. As the number of detected objects is explicitly present in the context, the QA model "simply" needs to reason over both sequences to predict the appropriate answer, as shown in the example of Figure 1. Because the QA model outputs the token position of the answer as well, it opens the possibility for users to understand where the language model focused its reasoning, and for us to produce the visual answer.

4 Experiments

Dataset. Our proposed Multi-Task prompt-RSVQA method is evaluated on the RSIVQA [36] dataset. This dataset contains remote sensing images and questions generated from the DOTA [12], SYDNEY [34], UCM [31], HRRSD [35], and AID [29] datasets. Accounting for a total of 108′898 samples, questions in the RSIVQA dataset can be grouped into four categories: *Counting* ("how many", numerical answer), e.g., Q: How many planes are there in this picture? A: 21; *Location* ("where", textual answer), e.g., Q: What is the location of the yellow car in this picture? A: left; *Presence* ("is there", yes/no answers), e.g., Q: Does this picture contain trees? A: yes; and *Scene* ("what is the scene", textual answer), e.g., Q: What is the theme of this picture? A: port. While the vast majority of the questions are generated automatically from templates, some samples are manually annotated. Inconsistencies in the answers have been noticed, for example numerical answers written as number and text (4 vs "four"), use of spacing at the end of the answer or spelling issues. Considering a classification strategy, these inconsistencies are problematic, as multiple classes can have the same meaning and thus confuse the model. Consequently, answers have been manually cleaned. For training and evaluation, data is divided into 80% training, 10% validation, and 10% testing as instructed in [36]. No exact lists of samples are proposed by the authors.

Experimental setup. The oriented-RCNN object detector is loaded with pre-trained weights on DOTA and HRSC2016 and run at inference. The Faster-RCNN is fine-tuned on HRRSD from pre-trained weights on COCO. We use a batch size of 16, with a learning rate of 10^{-4}, during 20 epochs. We consider detected objects with a probability equal or greater than 0.3, this value was experimentally determined. For the LC multi-label classification module, a ResNet-18 is fine-tuned on the multi-label version of the UCMerced dataset (from pre-trained weights on ImageNet [10]) with a batch size of 16, a learning rate of 10^{-4}, which stopped after 62 epochs. A threshold of 0.5 is used to consider a class present. We use three language models: For the question type classification, we use a distilBERT for sequence classification, fine-tuning it from the checkpoint *distilbert-base-uncased* [28] with a batch size of 64, a learning rate of 10^{-6} during 10 epochs. For the answer classification, we use a distilBERT from the pre-trained checkpoint *distilbert-base-uncased* [28] and fine-tune it on the non-numerical answers of RSIVQA (respectively with all answers for the ablation study without the QA module). The last model, roBERTa [16] QA, is used at inference only

(no fine-tuning). Two ablation studies are proposed: the answer prediction step is reduced to the classification head only (Figure 1(b)), and/or the image processing step is reduced to objects detection only (Figure 1(c)). RSVQA models [7, 19] are fine-tuned on the RSIVQA dataset for 20 epochs, with a 10^{-5} learning rate and a batch size of 20. The performances of the RSIVQA model are taken from [36] directly, and are accuracy scores only. As mentioned before, the exact train/validation/test sets being unpublished, comparisons with the RSIVQA model should be considered carefully.

Classification metrics evaluation. To evaluate the performance of our classification models, we use an overall accuracy and type-specific accuracy, allowing us to understand how well the model is performing on each question type. For *Counting* questions, the QA module predicts text as output, as opposed to a class, and thus cannot directly lead to an accuracy calculation. In order to still produce this metric for comparison, the textual answer is re-encoded into classes. However, these classes are never directly learned by this model head.

Regression metrics evaluation. As accuracy is ill-suited to evaluate performances of numerical predictions (i.e. penalizes identically all predictions not exactly equal to the targets), we use two common regression metrics: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), that are visually supported by "predictions vs. targets" scatter plots.

5 Results and discussion

We present the results first with the classification metrics, then with the regression metrics for counting samples, supported with scatter plots, and finally with the visual response our model can produce. Results for the question classification in step 3.2 (Section 3.2) are practically perfect due to the relative simplicity of the language vocabulary (not displayed). We compare our results against RSVQA models [7, 19, 36], as well as a zero-shot VQA model [25]. Finally, we analyse the importance of single modules with a series of ablation studies illustrated in Figure 1(b-c).

Accuracy scores ↑	Overall	Counting	Presence	Scene	Location
RSIVQA baseline [36]*	77.39	56.71	92.82	54.50	
RSVQA [19]	77.86	52.43	**95.60**	**55.43**	23.91
RSVQA-BERT-mutan [7]	77.89	52.52	95.53	54.84	50.00
Plug-and-Play VQA [25]	*51.22*	*24.78*	*75.43*	*0.51*	*0*
Ablation no LC, no QA	76.45	64.51	91.42	34.24	**52.94**
Ablation no LC	*69.00*	*39.11*	91.30	33.97	**52.94**
Ablation no QA	**79.52**	**64.75**	93.30	49.89	50.98
Multi-task prompt-RSVQA (our)	*72.19*	*39.15*	93.37	50.11	**52.94**

Table 1: Accuracy scores of our model against ablation studies, baseline architectures, and a zero-shot VQA model (Plug-and-Play). *Italics* are used to indicate ill-suited evaluation metrics, this is, accuracy for generative VQA and counting in our method. *The RSIVQA model [36] did not differentiate between *Scene* and *Location*.

Classification metrics results. Classification accuracy scores are summarized in Table 1. Our Multi-task prompt-RSVQA and its variants achieve competitive scores with the RSIVQA and RSVQA models, despite the semantic bottleneck (the image context) imposed in the architecture. This is for all categories except *Counting*. This is expected since counting questions are handled by the QA module (text extraction) instead of a classification model. The adequacy of classification metrics for this module as well as the Plug-and-Play VQA (PnP) needs to be discussed. PnP is directly transferred from VQA to RSVQA without any adaptations, and we test the zero-shot ability of this architecture on the RSIVQA dataset. The results do not meet the performances of the other models, especially for *Scene* and *Location*. PnP performs better on *Counting* and *Presence*, but is still well below classification models. As a generative model, the predictions consist of text instead of classes. While the answer prediction can be similar to the ground truth, it is often not exactly identical and thus evaluated negatively by classification metrics. Language generative models are rather evaluated with specific similarity metrics such as BLUE [20] or CIDEr [27]. As it would be irrelevant to use these metrics on the other RSVQA models performing a classification task, we still consider accuracy, but insist on handling comparisons carefully. The same argumentation holds in the case of text extraction performed by the QA module for numerical questions, motivating further evaluation in a regression setting (see Table 2 below). For reading ease, we used italics in Table 1 to highlight such edge cases. Our model outperforms all ablation versions for the other question types, showing that the separation of classification and regression heads boosts performance for these *Presence*, *Scene* and *Location* answer predictions.

Ablation studies. Diving into the ablation studies performed with our proposed model (bottom half of Table 1), we discuss the contribution of both the LC module in the image context, as well as the QA module for the answer prediction of samples of the *Counting* question type.

- Regarding the former, in particular for the *Scene* question type, and the *Presence* question type to a lesser extent, the LC context appears necessary. In these cases, distilBERT benefits from an image context containing information about land cover. This information helps characterizing classes spanning over large areas and unrelated to specific objects. In remote sensing, LC information can represent a large coverage of the image, containing very few objects, indistinguishable objects or completely lacking objects. Thus, this module is a necessary addition to detected objects for the description of remote sensing images.

- The ablations studies without a QA module consist of a classification-only model, treating all question types with the same head. In this case, the answer space also encodes numerical answers and is thus larger. In terms of *Presence*, *Scene* and *Location* samples, the accuracy scores are mostly similar to their corresponding multi-head variants. For *Counting* samples, the difference is important, which strongly impacts the overall scores as well. However, as explained, evaluating the text extraction performed by the QA module with an accuracy metric is not adequate as it does not learn the dataset classes (answers). Thus, rather than being compared to the Multi-task prompt-RSVQA, these ablations "no QA" should be considered as the classification-only equivalent of our model and used in comparison with other classification-only models. As such, it highlights the great improvement in *Counting*, 8 points compared to the RSIVQA baseline, of performing an explicit detection of objects.

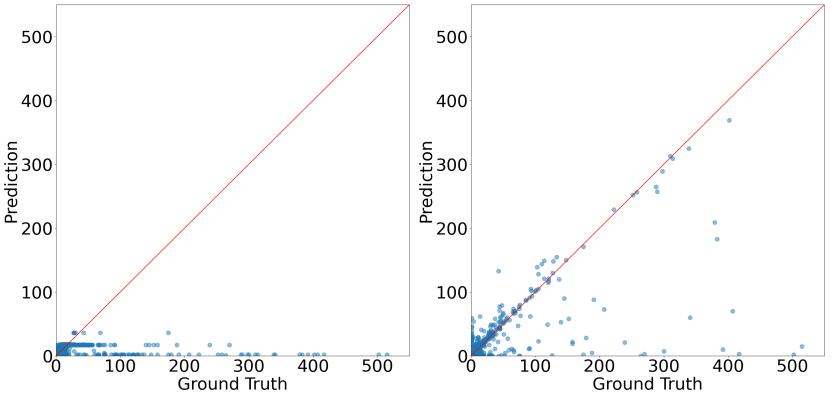

Figure 2: Scatter plots of counting predictions against ground truth for the classification-only *Ablation: no QA* (left) and our Multi-task prompt-RSVQA model with QA module (right).

Regression metrics results. Table 2 summarizes the error scores for numerical answers of *Counting* questions. In the case of our model, these answers are predicted by the RoBERTa QA module. As metrics, RMSE and MAE allow to go beyond the True/False evaluation of accuracy and inform on how close predictions are from ground truths as numerical values.

The best results, i.e. the smaller scores, are obtained with our model Multi-task prompt-RSVQA and its "Ablation no LC", both showing a considerable improvement on the other models. The identical scores calculated on the predictions of both experiments (4.26) highlight the fact that only the object detection part of the context is used by the QA module to retrieve/predict the answer, i.e. the presence of the LC description does not affect the results. As the QA module directly extracts the answer from the context, and thus the visual detector output, an improvement on this module would directly lead to an improvement of our model's final predictions.

Plug-and-Play VQA scores are the largest through all experiments (lower RMSE or MAE is better), which can be explained by two factors: the captions of image sub-parts generated inside the model's pipeline are not suitable to account for the number of elements over the entire image, and/or the captioning model struggles with the bird-eye viewpoint of remote sensing images and describing its content. Classification-only models (RSVQA baselines, as well as *Ablation no LC, no QA* and *Ablation no QA*) obtain relatively similar scores.

Regression metrics ↓	RMSE (Counting)	MAE (Counting)
RSVQA [19]	41.66	6.45
RSVQA-BERT-mutan [7]	40.94	6.24
Plug-and-Play VQA [25]	45.13	7.51
Ablation no LC, no QA	41.98	5.82
Ablation no LC	**35.42**	**4.26**
Ablation no QA	41.91	5.74
Multi-task prompt-RSVQA (our)	**35.42**	**4.26**

Table 2: Counting regression metrics results. Lower scores are better.

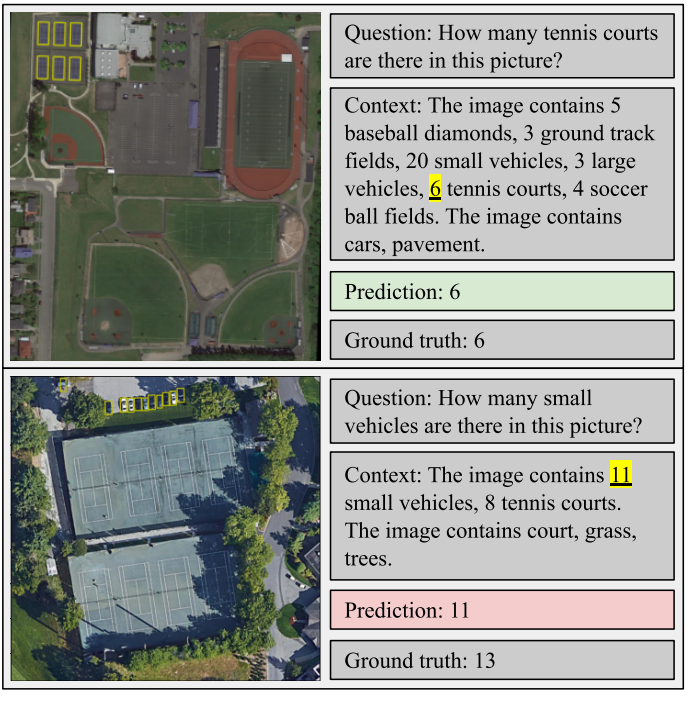

Figure 3: Two samples with visual answers using Multi-task prompt-RSVQA. In yellow, both the token position of the answer in the context and its spatial footprint in the image (bounding boxes) are highlighted.

MAE is distinctly better for the ablation studies, while RMSE is only slightly better for the RSVQA baselines. This indicates that the ablation studies are often closer to the targets, but RSVQA baselines tend to predict less outliers, as RMSE is more sensitive to them than MAE. Generally, a classification framework has the tendency to learn the most probable answer to a question or type of object, and thus will not adapt well to less conventional situations.

Figure 2 displays predictions against targets for *Ablation no QA* (left) and Multi-task prompt-RSVQA (right). In particular, it shows a better alignment of predictions obtained with the QA module (Multi-task prompt-RSVQA) around the perfect prediction line (red). While our Multi-task model predicts less frequently the exact counting answer (as seen in accuracy scores), it is consistently closer to the ground truth. With our model, we argue that closer predictions give a better information to users than learned frequent classes of a dataset. The explicit detection of the image content makes the task easier and more transparent for the language model and leads to better performances. Moreover, it creates a semantic bottleneck, the context, useful for RSVQA users to interpret the predictions.

Visual answer. Using explicit object detection along with a QA module for the answer prediction allows to return a visual answer in addition to text for *Counting* questions. The QA module outputs the token position of the extracted answer in the context sentence. Using this information, the object class is identified and their bounding boxes can be plotted on the input image as exemplified in Figure 3.

The visual answer is a valuable and desirable feature for practitioners. Indeed, if we were interested in analyzing images after a natural disaster, a valid question for the model would be "How many damaged buildings are in this picture?". The answer would be even more insightful if we could localize these buildings in the image.

6 Conclusion

We proposed a Multi-task prompt-RSVQA model based on describing the image content (context) and predicting answers with classification or question answering modules, depending on the question type. As shown in our experiments, our model improves performances on *Counting* questions with numerically closer and more reliable predictions with respect to the ground truth. This improvement is possible thanks to the explicit detection of the image content, in particular the presence of objects, and the use of a question answering (QA) module for the answer prediction. However, moving away from classification to text extraction of the answer requires to rethink the evaluation metrics. Thus, for numerical answers, we use common regression metrics RMSE and MAE to evaluate, along with scatter plots to illustrate predictions against targets. As a modular model, improvements in the separate modules, in particular the visual ones, is key for improving performances, thus mitigating the relative scarcity of annotated data available for large training in RSVQA. Finally, our strategy aims towards more interpretability in RSVQA and allows to produce a visual answer, grounding the prediction in the image. Considering the motivations and potential users of RSVQA, this represents a novel and highly relevant step for the task.

References

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural Module Networks. In *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, Las Vegas, NV, USA, 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.12. URL http://ieeexplore.ieee.org/document/7780381/.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the 2015 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile, 2015. IEEE. URL https://openaccess.thecvf.com/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html.

[3] Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Mohamed Lamine Mekhalfi, Mansour Abdulaziz Al Zuair, and Farid Melgani. Bi-Modal Transformer-Based Approach for Visual Question Answering in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, July 2022. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2022.3192460. URL https://ieeexplore.ieee.org/document/9832935/.

[4] Prajjwal Bhargava and Vincent Ng. Commonsense Knowledge Reasoning and Generation with Pre-trained Language Models: A Survey. In *Proceedings of the 2022 AAAI Conference on Artificial Intelligence*, volume 36, pages 12317–12325, Virtual, 2022. AAAI. doi: 10.1609/aaai.v36i11.21496. URL https://ojs.aaai.org/index.php/AAAI/article/view/21496.

[5] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text Retrieval: A Survey on Recent Research and Development. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5410–5417, Vienna, Austria, 2022. IJCAI. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/759. URL https://www.ijcai.org/proceedings/2022/759.

[6] Christel Chappuis, Sylvain Lobry, Benjamin Kellenberger, Bertrand Le Saux, and Devis Tuia. How to find a good image-text embedding for remote sensing visual question answering? In *In Proceedings of MACLEAN: MAChine Learning for EArth ObservatioN Workshop, co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2021)*, Virtual, 2021. CEUR-WS. URL http://arxiv.org/abs/2109.11848. arXiv: 2109.11848.

[7] Christel Chappuis, Vincent Mendez, Eliot Walt, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Language Transformers for Remote Sensing Visual Question Answering. In *Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2022)*, pages 4855–4858, Kuala Lumpur, Malaysia, 2022. IEEE. ISBN 978-1-66542-792-0. doi: 10.1109/IGARSS46834.2022.9884036. URL https://ieeexplore.ieee.org/document/9884036/.

[8] Christel Chappuis, Valérie Zermatten, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Prompt-RSVQA: Prompting Visual Context to a Language Model for Remote Sensing Visual Question Answering. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1372–1381, New Orleans, LA, USA, 2022. IEEE. URL https://openaccess.thecvf.com/content/CVPR2022W/EarthVision/html/Chappuis_Prompt-RSVQA_Prompting_Visual_Context_to_a_Language_Model_for_Remote_CVPRW_2022_paper.html.

[9] Bindita Chaudhuri, Begum Demir, Subhasis Chaudhuri, and Lorenzo Bruzzone. Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):1144–1158, October 2017. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2017.2760909. URL http://ieeexplore.ieee.org/document/8089668/.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Miami, FL, USA, 2009. IEEE. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848/.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, MN, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL http://aclweb.org/anthology/N19-1423.

[12] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7778–7796, October 2021. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2021.3117983. URL https://ieeexplore.ieee.org/document/9560031/.

[13] Rafael Felix, Boris Repasky, Samuel Hodge, Reza Zolfaghari, Ehsan Abbasnejad, and Jamie Sherrah. Cross-Modal Visual Question Answering for Remote Sensing Data. In *Proceedings of the 2021 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–9, Gold Coast, Australia, 2021. IEEE. ISBN 978-1-66541-709-9. doi: 10.1109/DICTA52665.2021.9647287. URL https://ieeexplore.ieee.org/document/9647287/.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, 2016. IEEE. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the 2014 European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich, Switzerland, 2014. Springer International Publishing. ISBN 978-3-319-10601-4 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48. URL http://link.springer.com/10.1007/978-3-319-10602-1_48. Series Title: Lecture Notes in Computer Science.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

[17] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A High Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines:. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, pages 324–331, Porto, Portugal, 2017. SciTePress. ISBN 978-989-758-222-6. doi: 10.5220/

0006120603240331. URL http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006120603240331.

[18] Sylvain Lobry, Diego Marcos, Benjamin Kellenberger, and D. Tuia. Better Generic Objects Counting when Asking Questions to Images: A Multitask Approach for Remote Sensing Visual Question Answering. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2020:1021–1027, August 2020. ISSN 2194-9050. doi: 10.5194/isprs-annals-V-2-2020-1021-2020. URL https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/V-2-2020/1021/2020/.

[19] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8555–8566, May 2020. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2020.2988782. URL https://ieeexplore.ieee.org/document/9088993/.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. URL https://dl.acm.org/doi/pdf/10.3115/1073083.1073135.

[21] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding. *IEEE Access*, 9:89644–89654, June 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3090981. URL https://ieeexplore.ieee.org/document/9460988/.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015)*, volume 28, pages 91–99, Montreal, QC, Canada, 2015. MIT Press. ISBN 978-1-5108-2502-4. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019. ISBN 978-1-66542-418-9. URL http://arxiv.org/abs/1910.01108. arXiv:1910.01108 [cs].

[24] Tim Siebert, Kai Norman Clasen, Mahdyar Ravanbakhsh, and Begüm Demir. Multi-modal fusion transformer for visual question answering in remote sensing. In *Proceedings of the Image and Signal Processing for Remote Sensing XXVIII*, page 21, Berlin, Germany, 2022. SPIE. ISBN 978-1-5106-5537-9 978-1-5106-5538-6. doi: 10.1117/12.2636276. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12267/2636276/Multi-modal-fusion-transformer-for-visual-question-answering-in-rem10.1117/12.2636276.full.

[25] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 951–967, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.67.

[26] D. Tuia, R. Roscher, J. D. Wegner, N. Jacobs, X. X. Zhu, and G. Camps-Valls. Towards a collective agenda on AI for earth science data analysis. *IEEE Geosci. Remote Sens. Mag.*, 9(2): 88–104, 2021. URL https://ieeexplore.ieee.org/document/9456877.

[27] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.html.

[28] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

[29] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, April 2017. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2017.2685945. URL http://ieeexplore.ieee.org/document/7907303/.

[30] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for Object Detection. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3500–3509, Montreal, QC, Canada, 2021. IEEE. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.00350. URL https://ieeexplore.ieee.org/document/9710901/.

[31] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, San Jose, California, USA, 2010. ACM. ISBN 978-1-4503-0428-3. doi: 10.1145/1869790.1869829. URL https://dl.acm.org/doi/10.1145/1869790.1869829.

[32] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research*, August 2022. URL https://openreview.net/forum?id=Ee277P3AYC.

[33] Zhenghang Yuan, Lichao Mou, Qi Wang, and Xiao Xiang Zhu. From Easy to Hard: Learning Language-guided Curriculum for Visual Question Answering on Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, May 2022. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2022.3173811. URL http://arxiv.org/abs/2205.03147. arXiv:2205.03147 [cs].

[34] Fan Zhang, Bo Du, and Liangpei Zhang. Saliency-Guided Unsupervised Feature Learning for Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4):2175–2184, September 2014. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2014.2357078. URL http://ieeexplore.ieee.org/document/6910306/.

[35] Yuanlin Zhang, Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu. Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5535–5548, March 2019. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2019.2900302. URL https://ieeexplore.ieee.org/document/8676107/.

[36] Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, May 2021. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2021.3079918. URL https://ieeexplore.ieee.org/document/9444570/.