Prompt–RSVQA: Prompting visual context to a language model for Remote Sensing Visual Question Answering

Christel Chappuis
EPFL
Sion, Switzerland
christel.chappuis@epfl.ch

Valérie Zermatten
EPFL
Sion, Switzerland
valerie.zermatten@epfl.ch

Sylvain Lobry
Université Paris Cité
F-75006 Paris, France
sylvain.lobry@u-paris.fr

Bertrand Le Saux
European Space Agency Φlab
Frascati, Italy
bertrand.le.saux@esa.int

Devis Tuia
EPFL
Sion, Switzerland
devis.tuia@epfl.ch

Abstract

Remote sensing visual question answering (RSVQA) was recently proposed with the aim of interfacing natural language and vision to ease the access of information contained in Earth Observation data for a wide audience, which is granted by simple questions in natural language. The traditional vision/language interface is an embedding obtained by fusing features from two deep models, one processing the image and another the question. Despite the success of early VQA models, it remains difficult to control the adequacy of the visual information extracted by its deep model, which should act as a context regularizing the work of the language model. We propose to extract this context information with a visual model, convert it to text and inject it, i.e. prompt it, into a language model. The language model is therefore responsible to process the question with the visual context, and extract features, which are useful to find the answer. We study the effect of prompting with respect to a black-box visual extractor and discuss the importance of training a visual model producing accurate context.

1. Introduction

Despite its potential, Earth observation (EO)-based information still remains difficult to access, mostly because of the technical requirements needed to convert the raw image data into actionable information (including the limited availability of vast labeled sets and the need for advanced machine learning skills). New ways to extract relevant information from images bypassing those requirements are needed to unleash the full potential of EO [10, 32] for the benefit of various application fields, such as environmental monitoring, agriculture, urban planning, tourism, etc.

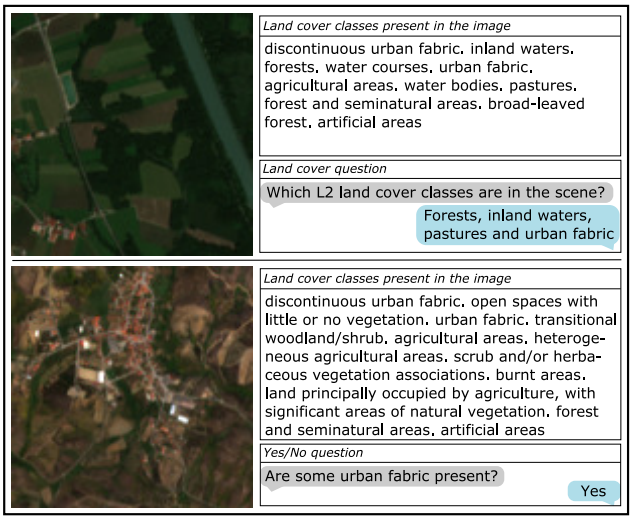

Figure 1. RSVQAxBEN triplets samples with visual ground truth.

Visual Question Answering (VQA) [2] was introduced as a generic and user-friendly way to interact with image-based products, fostering the transition from arbitrary and controlled tasks (e.g., image classification) to diverse applications accessible to a wide audience. In a nutshell: given an image and a question about it in natural language, the VQA task consists in predicting a text answer. Remote sensing visual question answering (RSVQA) [20] followed this principle to enable a wider use of remote sensing images. A couple of examples from [19] are displayed in Figure 1.

Current RSVQA methodologies rely on a joint latent representation of visual and textual information, obtained with fusion and/or attention mechanisms, to derive the an-

swer. When considered in isolation, the visual and language tasks can benefit from the specific advances from computer vision and natural language processing (NLP). Foundation models [3] such as the BERT [8] and GPT-3 [4] families or G-Shard [17] accelerated NLP significantly. However, constructing a meaningful bi-modal latent space from scratch requires considerable efforts. Huge computational resources are needed for the re-training of these pre-trained models on specific, application-oriented tasks.

Prompt-based inference [4] was introduced as a new paradigm for leveraging these huge models. Instead of re-training from end to end a large language model, *prompting* adds a few keywords to the input text. These keywords act as a contextual guidance and make predictions with no or light fine-tuning only. In VQA, instead of relying on a latent, thus abstract, joint representation of visual and textual information, the additional keywords can be cast as the context given by the image to the language model. The abstract visual features extractor is replaced with a keywords generator, which produces a textual description of the image, then used in the language model. In other words, the remote sensing image, translated into a textual representation, can be used as prompt (also referred to as context in this paper) for the question in a language-only model.

Relying only on a language model, prompted with context extracted by visual analysis of the image, bears two important advantages: first, it creates an interpretable bottleneck (the keywords generator) that can be used to understand which elements of the image are used to answer the question; second, it places language as the fundamental reference modality to translate other modalities to. While our interest in this paper is a visual modality (the remote sensing image), other sources of information and/or knowledge could be added to the prompt by simply converting them into a set of insightful keywords for the language model. Translating modalities to a reference could be computationally more interesting than defining a new appropriate latent multi-modal space for every new situation encountered.

In this study, we investigate an alternative method to the bi-modal representation for remote sensing VQA [9, 20, 41] based on image-based keywords generation, or prompting. We refer to it as *Prompt–RSVQA*, which stands for *Prompt* *R*emote *S*ensing *V*isual *Q*uestion *A*nswering. The novel architecture places language in the center as the reference modality, and translates the visual information into words to guide the language model. An attention-based Transformer language model attends both the question and description of the image to produce an answer. Through wide experimentation on the RSVQAxBEN dataset [19] (of which samples are showed in Figure 1), we aim at assessing each modality individually and at understanding the influence of visual predictions on the subsequent language model. Using prompting, we obtain answers that are i) more accurate, ii) more interpretable, since derived from humanly readable visual elements and, iii) even when incorrect, still semantically closer to the right answer than a model based on an abstract bi-modal latent representation.

After reviewing recent advances on VQA and prompt-based methods in Section 2, we present our method in Section 3. The experimental procedure is detailed in Section 4 and we present and discuss our results in Section 5.

2. Related work

While the initial methodology [2] of **Visual Question Answering** relies on extracting features from both modalities, combining them in a fusion whose product is classified into an answer, methods aiming at expliciting the interaction between vision and language have been more and more researched: attention mechanisms have rapidly been introduced to guide the search of meaningful content across modalities [1, 38]. Graph representations are suggested to enhance the joint reasoning across modalities [30]. Recently, the VQA task along with multiple other vision-and-language problems are tackled using large-scale pre-training of networks on text-image datasets [6, 42].

The last years have witnessed a radical switch toward the usage of large text-image models, also referred to as **foundation models**, such as CLIP [23], UNITER [6], MERLOT [42], Florence [39] or UFO [35]. These models have demonstrated their strength on multiple vision-and-language tasks, leveraging colossal quantities of data in unsupervised pre-training. Modern foundation language models involve hundreds of billions of learnable parameters (GTP-3 [4], 175B; LaMDA [31], 137B; Megatron-Turing [27], 530B). While their performances are undeniable, these models come with huge computations cost, even for re-training. Following the publication of GPT-3, the concept of **language prompting** raised an increasing interest [18]. Its core principle is to manipulate input data instead of the model parameters. While the initial idea is inference-based, it was transferred to smaller language models such as the BERT family with light fine-tuning [11, 25]. Language prompts have been applied to the task of question answering [45] and VQA [14, 37].

Remote sensing visual question answering was initially proposed by [20] with a baseline methodology and a dataset in which questions and answers are derived from OpenStreetMap using the CLEVR protocol [15]. Contributing to the research field, other datasets have been published [40, 44] while methodologies implementing different fusion mechanisms [5], attention mechanisms [44], curriculum learning [41] or object detectors [9] have been experimented with. A new dataset "RSVQA meets BigEarthNet" (RSVQAxBEN [19]) focuses on questions/answers relative to land cover and is used in this work. The ground truth answers are generated solely from an easily accessible visual

ground-truth, BigEarthNet labels [29].

In this paper, we propose to use discrete tokens from the land-cover vocabulary to prompt a language model from the BERT family. By doing so, we consider the visual information extracted from the remote sensing image as a context information guiding the language model in the RSVQA task. To the best of our knowledge, this is the first utilisation of the concept of prompting in remote sensing.

3. Proposed method *Prompt–RSVQA*

Figure 2 summarizes our proposed methodology. We first use a visual model (Section 3.1) to predict land cover classes present in the image and convert them into text in a context (Section 3.2). These classes and the question are then passed to the language model, and the vector produced is classified into an answer using a MLP (Section 3.3).

3.1. Visual model

The objective of the visual model is to predict the land cover classes that are present in the RGB images. In this work, we use a ResNet-50 [12] model pre-trained on ImageNet [7]. ResNet-50 [12] is a CNN image classifier that uses skip connections to reduce the impact of the vanishing gradients problem. The original ResNet-50 architecture is adapted to the multi-label task by replacing the final softmax layer by a sigmoid layer as the activation function. As a result, the model provides a presence score ranging from 0 to 1 for each land cover class.

3.2. Context construction

The output of the ResNet-50 visual model is transformed into a context that serves as an input to the language model. First, the predictions from the visual output are thresholded (with a threshold θ), giving a one-hot vector indicating the presence of each considered land cover class. The labels of the selected predictions are then retrieved and concatenated together to form a sequence of words. Each label is separated by a full stop character, as shown in the example below, where *'class2'* and *'class4'* would be textual names of the classes (e.g. *'marine waters'*, *'forest'*, etc).

$$\begin{bmatrix}0.3\\0.7\\0.2\\0.9\end{bmatrix} \xrightarrow{\theta} \begin{bmatrix}0\\1\\0\\1\end{bmatrix} \rightarrow \begin{bmatrix}\text{`}class2\text{'} & . & \text{`}class4\text{'}\end{bmatrix}$$

3.3. Language model

The language model takes the question and the textual representation of the image as inputs to produce a feature vector jointly representing both visual and textual information. DistilBERT [24] is used as the language model in our framework. DistilBERT is an attention-based Transformer [33] architecture. It is a light, distilled version of the Bidirectional Encoder Representations from Transformers (BERT) [8] pre-trained on the BookCorpus [46] dataset and the English Wikipedia. Compared to BERT, it is shallower by a factor of 2 (6 self-attention layers in distilBERT against 12 for BERT). DistilBERT is trained using knowledge distillation [13], retaining up to 97% of BERT base performance on GLUE benchmark [34].

The two input sequences, question and visual context, are tokenized. Each word is transformed into a number, or token. The input vector contains first the list of tokens for the question, then a special separation token, followed by the list of tokens corresponding to the context. The tokenization process is followed by distilBERT language model. In the model, tokens along with their position and type are embedded in its space before being passed to the encoder, i.e. the six attention layers. The output of the language model is a vector of dimension 768.

Finally, the answer prediction is framed as a classification task, where $1'000$ classes represent a set of pre-defined answers. The output of the language model is projected to the answer space with a MLP (one hidden layer of size 256).

4. Data and setup

4.1. Data

Experiments are conducted on the large-scale dataset RSVQAxBEN [19] that focuses on land cover questions/answers. This allows us to fully supervise both models for an initial exploration of Prompt-RSVQA. This dataset is derived from the BigEarthNet dataset [29].

BigEarthNet (BEN) [29] is a large-scale benchmark dataset for multi-label land cover classification. It contains $590'326$ Sentinel-2 image patches collected in 2017 and 2018 over ten countries in Europe. The images have a spatial extent of 1.2km$\times$1.2km and a spatial resolution ranging from 10 to 60m for the 13 spectral bands. Images with significant cloud cover, cloud shadows or covered by seasonal snow were discarded in the dataset construction.

The original labels are derived from the CORINE Land Cover (CLC) inventory. The CLC established a hierarchy of land cover classes with 3 levels of labels that are split into 5 coarse (L1), 15 intermediate (L2) and 43 fine-grained (L3) land cover categories. A few classes with similar designations at different hierarchy levels (e.g. water bodies, pastures) are counted as a single label leading to a total of 61 classes. The accuracy of this reference data is estimated to be over 92% overall, and of 87% when considering only L3 classes [22]. BigEarthNet associates each image with the L3 CLC labels present in the spatial extent of the patch. The data split used in [29] between training, validation and testing sets is done randomly.

RSVQAxBEN dataset [19] is derived from the 10m res-

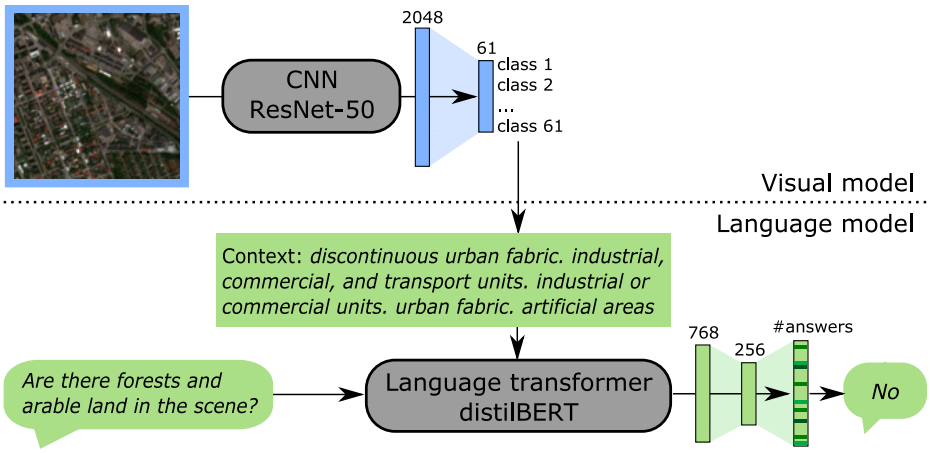

Figure 2. Our proposed method *Prompt–RSVQA*, translating the image into a context and then using it in a language-only model.

olution RGB images and the labels from BEN. Question/answer pairs are constructed from the CLC labels (L1 and L2 are derived from the L3 labels provided by BEN) using a stochastic procedure (Figure 1). Because there is no ambiguity in the construction procedure, it is to be noted that questions can be answered perfectly given the correct list of CLC classes present in the image. There are two types of questions: either *Yes/No* (e.g. "Is there a forest or a water area in this image?") or *Land cover* (e.g. "Besides natural vegetation, which land cover classes are present?") questions. Retrieving an extra information, i.e. the CLC labels, in addition to the image/question/answer VQA triplets, offers the opportunity to monitor the visual part of the pipeline, and to obtain an interpretable semantic bottleneck [21]. Moreover, it allows to study the upper bound on accuracy, i.e. the situation where perfect CLC labels are used as context information.

This dataset splits in training, validation and testing sets differs from the ones in BigEarthNet. It separates the samples by the latitude coordinate of the images, rather than randomly. This splitting methodology induces a shift in categories distribution between training/validation/testing sets since they come from geographically distant areas. While avoiding spatial correlation between training and evaluation, this domain shift also makes the land cover classification task more challenging. In this work, we use this split.

Finally, in RSVQAxBEN, there are a total of $28'049$ possible answers. However, a procedure, following the work in [19], is used to select the 1'000 most frequent answers (of the training set). These answers cover respectively 98.1%, 99.2% and 98.9% of the answer spaces for the training, validation and testing sets.

4.2. Model training procedure

The visual model and the language model (including the answer prediction) are fine-tuned separately in this study. The first considers the image alone and predicts the CLC labels (at the L1, L2 and L3 levels). The second is fine-tuned by using a perfect visual prediction, i.e. using directly the L3 labels from BigEarthNet and the corresponding L1 and L2 labels as the (perfect) context for answering the question. Learning the model in an end-to-end fashion is not straightforward, since it requires to back-propagate through the prompting operation, and is left for future research.

As stated above, fine-tuning the language model with an exact context allows to assess an upper bound on performances for the complete framework. This model is referred to as *visual oracle model* in the present study. On the contrary, a *visual blind model*, considering only the question as input, without any form of information from the image, is trained to assess a lower bound of performances to be expected on the RSVQAxBEN dataset. The method *Prompt–RSVQA* is displayed in Figure 2, while the *visual oracle* and *visual blind* models are illustrated in Figure 3.

The visual model is fine-tuned with the Adam optimizer [16] for 100 epochs, with an initial learning rate of 10^{-4}, reduced by a factor 0.1 every 30 epochs. We use a batch size of 64. Random vertical and horizontal flips, as well as random 90 degrees rotations, are performed during training as data augmentation strategies. We use the binary cross entropy loss that combines a sigmoid function with the cross entropy loss to have a multi-label output where each class receives a scores between 0 and 1. The number of output classes is fixed to 61, which corresponds to the ensemble of L1, L2 and L3 classes in the CLC hierarchy. We observe that training the model with all three hierarchical levels of land cover slightly improves the model performances in comparison to training it with the L3 labels only.

The language model includes the language Transformer and the classification layers to predict the answers. The Adam optimizer [16] is used here as well. The model is

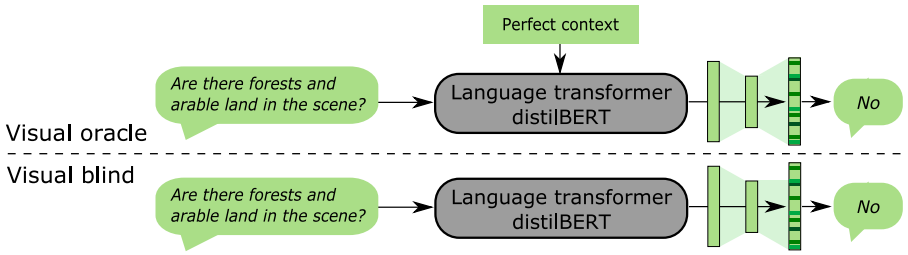

Figure 3. The *visual oracle* and *visual blind* models; While the first benefits from a perfect visual context that is the ground-truth of BigEarthNet image data, the second never sees the image.

trained for 10 epochs with a batch size of 100. The loss is computed with cross entropy. The "transformers" package from the open library Huggingface provides an implementation of both the specific tokenizer and language model. As tokeniker, we use DistilBertTokenizer (30'000 tokens vocabulary). This is identical to BertTokenizer, which is based on the subword tokenization algorithm WordPiece [26, 36]. The pre-trained weights of the distilBERT language model are retrieved from the common checkpoint 'distilbert-base-uncased'. DistilBERT (excluding the embedding part) is fine-tuned with a learning rate of 10^{-6}. The classification layers are trained with a learning rate of 10^{-5}.

4.3. Performance evaluation

In our results, evaluations are divided in two parts: those referring to our final task VQA and those referring to the bottleneck task, the prediction of CLC classes.

4.3.1 Multilabel CLC classification

Performance evaluation for multi-label classification is more complex than for traditional classification since they require specific metrics that also consider partially correct answer, instead of a binary correct/incorrect evaluation.

The partial correctness of classification is evaluated with the **F1-score** that computes a harmonic mean of the precision and recall. We use the micro F1-score to compare predictions and ground truth globally on every categories.

The **exact match ratio (MR)** computes the fraction of correctly classified samples, i.e. the samples whose predicted labels exactly correspond to the ground truth labels. This metric is similar to accuracy for traditional classification. For N samples, it is expressed as follow [43]:

$$MR = \frac{1}{N}\sum_{i=1}^{N} I(\mathbf{y}_i = \hat{\mathbf{y}}_i), \tag{1}$$

where $\mathbf{y}_i$ is a vector containing the ground truth labels, $\hat{\mathbf{y}}_i$ the predicted labels expressed as a one-hot encoded vector of land cover classes, and I the indicator function (returning 1 if the condition is true, 0 otherwise).

To cope with this issue and provide a sense of *how wrong* a prediction is, we report the **Hamming distance (HD)** between the prediction vector and the ground truth. It corresponds to the number of times, on average or per sample, an occurrence of class label is incorrectly predicted [28]. In other words, it represents the number of binary labels to modify in the prediction to obtain the ground truth, equally considering missed and wrongly predicted labels. It can be formulated as follow:

$$HD = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} I(y_{ij} = \hat{y}_{ij}), \tag{2}$$

where y_{ij} and $\hat{y}_{ij}$ are the binary class target and prediction, for C land cover categories. The smaller the HD, the better the performance of the model, thus a HD of 0 indicates a perfect prediction. This example shows the difference between the per sample MR and HD scores when comparing two incorrect prediction vectors $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2$ with target $\mathbf{y}$:

$$\begin{aligned} \mathbf{y} &= \begin{bmatrix}0 & 1 & 1 & 1 & 0\end{bmatrix}; \\ \hat{\mathbf{y}}_1 &= \begin{bmatrix}0 & 0 & 1 & 1 & 0\end{bmatrix}; \\ \hat{\mathbf{y}}_2 &= \begin{bmatrix}1 & 0 & 0 & 1 & 0\end{bmatrix}; \\ \rightarrow MR &= \begin{bmatrix}0\\0\end{bmatrix}, \quad HD = \begin{bmatrix}1\\3\end{bmatrix}. \end{aligned}$$

4.3.2 VQA downstream task

VQA is mostly studied in terms of **accuracy of the answers** provided. In the experiments, we report accuracy figures globally (over all samples) and per type of question ("Yes/No" or "Land cover" subsets). With the accuracy metric, all mistakes count the same, regardless of their similarity to the ground truth. Thus, we additionally analyse the HD (Eq. (2)) between the answers provided; HD allows to understand if the predicted answers are critical mistakes (several classes are wrong or missing) or more nuanced ones (only one CLC class is incorrect in the answer). To compute the HD on the VQA answers, we convert each potential answer into a 64-dimensional one-hot vector, $y_{VQA} \in \{0,1\}^{64}$, with the binary entries ordered as follows: *'Yes'*,*'No'*,*'None'*, CLC classes.

Method	Classes	F1	HD	MR
Our method	L1, L2, L3	0.746	3.4	15.6%
Our method	L3	0.682	1.9	18.5%
S-CNN-RGB [29]		0.676	-	-

Table 1. Multi-label classification results with the F1-score, Hamming distance (HD) and exact match ratio (MR). The results are displayed for our method on all the categories (61), and on the L3 categories (43) only, to allow for a comparison with S-CNN-RGB [29] that predicts only L3 classes.

Method	Global	Yes/No	Land cover
Visual oracle	98.81%	99.90%	93.79%
Visual blind	65.36%	75.85%	17.30%
RSVQA [19]	69.83%	79.92%	20.57%
Prompt–RSVQA (ours)	**75.40%**	**86.07%**	**26.56%**

Table 2. Results in accuracy, comparing performances from [19] with *visual oracle*, *visual blind*, and our method *Prompt–RSVQA*. "Blind" and "oracle" results give, respectively, a lower and upper bounds of potential results on the RSVQAxBEN dataset.

5. Results and discussion

In this section, we present results, first on the visual model only (predicting CLC classes, Section 5.1), then for the VQA task (Section 5.2). Finally, we discuss the robustness of the model to perturbations in the CLC results (Section 5.3) and how critical mistakes are with respect to the target (Section 5.4).

5.1. Visual model

First, we discuss the results of our visual classifier predicting CLC classes displayed in Table 1. Our resulting F1-score on the L3 categories is similar to the published one in [29], even though our dataset split is more challenging due to the distribution shift between the train, test and validation sets (see Section 4.1). On average, the visual model makes 3.4 mistakes on the 61 classes, among which 1.9 are erroneous predictions on L3 categories, as indicated by the Hamming distance.

5.2. Language model and the VQA task

In this section, VQA results on the test set for the *visual oracle model*, *visual blind model*, and proposed method are compared against the baseline method established in [19]. Performances in accuracy are displayed in Table 2.

Upper and lower bounds. First, we analyse the results of our upper and lower bounds, the *visual oracle* and the *visual blind* models, respectively. As a reminder, the former has perfect knowledge of the CLC classes for prompting, while the latter does not use any form of visual information in any way and relies on biases in the questions / answers distributions to predict the answer.

- *Visual oracle*. Fine-tuning the distilBERT language model with a perfect input from the visual part (using BigEarthNet labels) allows to define the upper bound of performances. As displayed in Table 2, the perfect CLC input used as prompt leads the language model and subsequent classification layers to near-perfect VQA predictions. In fact, considering the way the answer space is restricted, as described in Section 4.1, the best attainable global accuracy on the testing set is 98.9%. The *visual oracle model* is thus only at about 0.1% point from what is truly achievable in this situation. With this result, we can see that a light Transformer like distilBERT manages to solve the language problem of the RSVQAxBEN dataset. This is relatively unsurprising, especially considering the templates-based procedure followed to build the questions. While the use of large language models may be argued at this point, we believe it will become necessary with the diversification of topics and the increasing complexity of language syntax used.

- *Visual blind model*. In the closed setting of a dataset, each question has a limited number of possible answers. For "Yes/No" questions, the answer has in fact only two options ("yes" or "no") and if the distribution is balanced, randomly picking either of the two options will lead to a 50% accuracy. However, the distribution is often imbalanced and thus blindly selecting the most common answer will lead to a reasonably good accuracy without even checking in the image. As shown in Table 2, the accuracy of the *visual blind model* meets this expectations, achieving impressive performances, only 3-4 points below the performances of [19]. We strongly feel the importance of raising awareness about biases in remote sensing VQA datasets and defining a lower bound performance to illustrate it.

Our *Prompt–RSVQA* method outperforms by about 6% the global accuracy and question type specific accuracy of RSVQA [19]. It outperforms by about 10% the *visual blind model*, but considerable progress is still needed to reach the performances of the *visual oracle model*, in particular regarding questions related to land cover classes (26.56% accuracy). This is unsurprising, since they are typically more complex than Yes/No questions (86.07% accuracy).

Performances of the downstream VQA task are dependent on the threshold applied to the CLC predictions. The search for the best θ is conducted on the validation set of

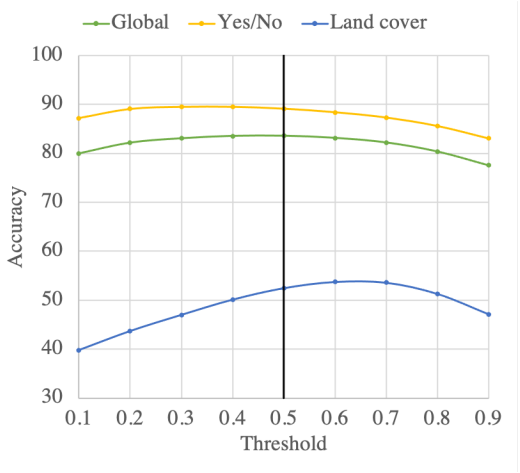

Figure 4. The threshold selection of the visual output is performed on the validation set of RSVQAxBEN, using the global accuracy of the RSVQA task, and represented with a black vertical line.

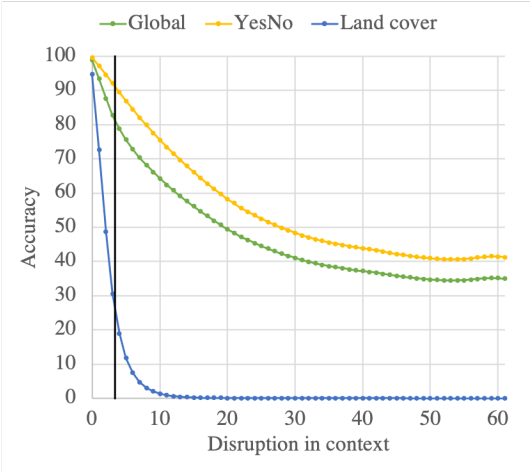

Figure 5. Disruption of the fine-tuned language model at inference, on the validation set. The black vertical line is the actual average error rate of our visual model (HD 3.4, Table 1).

RSVQAxBEN and is illustred in Figure 4. The best threshold, $\theta = 0.5$, is chosen for the highest global accuracy. Interestingly, it appears the more appropriate threshold for each question type differs. While the value best suited for the "Yes/No" questions is slightly lower, between 0.3 and 0.4, the best value for the "Land cover" questions lies between 0.6 and 0.7. In other words, the threshold needs to be more restrictive, i.e. select fewer labels, for the "Land cover" questions than for "Yes/No".

Comparing the vertical axis in Figure 4 with the results in Table 2, there is a gap in performances between validation and testing results. As described in Section 4.1, the strategy to split the dataset into training, validation and testing sets in RSVQAxBEN imposes different distributions between the sets. This challenge likely explains the gap in performances between validation and testing sets.

5.3. Sensitivity to the visual predictions

While the light fine-tuning of the language model is performed with a synthetic perfect input from the visual part, the real performances of the visual model are not flawless (Table 1). To better understand the quality of results required from the visual part to perform well, a perturbation analysis is conducted during inference on the language part (Figure 5). The perfect input used for prompting is corrupted randomly by adding/removing CLC classes. Considering each possible label in a binary manner, activated or not, class entries are swapped (i.e. in the prediction vector, zeroes are changed to ones and vice versa). From 0 to 61 disruptions in each prediction vector are performed to assess the influence on the final task of answer prediction.

As expected, disruptions of the visual input affect more strongly the "Land cover" question type. Predicting erroneously more than 4 labels from the image makes the accuracy for this question type drop below 20%, i.e. below the performance reported by [19]. Worsening to 10 mistakes or more pushes this accuracy to nearly 0%. On the other hand, the loss in performance on the "Yes/No" questions is less dramatic. Its accuracy decreases from almost 100% to about 40% when most classes responses have been perturbed. The number of samples of this question type represent about 80% of the dataset, influencing more heavily the global results, as it clearly appears in Figure 5.

5.4. Does our method lead to better mistakes?

Our final analysis relates to the way the answer are constructed in RSVQAxBEN. As its name suggests, the "Yes/No" questions are answered by "Yes" or "No", while the "Land cover" questions can be answered by "None" or a sequence of one or more classes of land cover. Given 61 different land cover classes, the number of possible combination of varying length is extremely large. This explains the motivation for limiting the answer space to the $1'000$ most common answers, instead of the $28'049$ full labels set variety in the dataset, a lot of which occur only once.

However, some mistakes are more severe than others: missing a single CLC class is definitely less critical than incorrectly predicting five, or answering 'Yes' when asked which CLC classes are present in the image. To study whether our strategy leads to less critical mistakes, we consider the Hamming Distance (HD). Average results are displayed in Table 3. The distribution of HD for each studied case is drawn in Figure 6. A distance of 0 indicates the prediction matches exactly the target.

As expected, the average global distance in Table 3 is best for the *visual oracle model* and worst for the *visual*

Method	Global	Yes/No	Land cover
Visual oracle	0.07	0.00	0.40
Visual blind	0.89	0.48	2.73
RSVQA [19]	0.81	0.40	2.80
Prompt–RSVQA (ours)	**0.59**	**0.28**	**2.02**

Table 3. Hamming distances in answer post-processing, comparing our method *Prompt–RSVQA* with [19] and *visual blind*. *Visual oracle* gives ideal performances with a perfect context classifier.

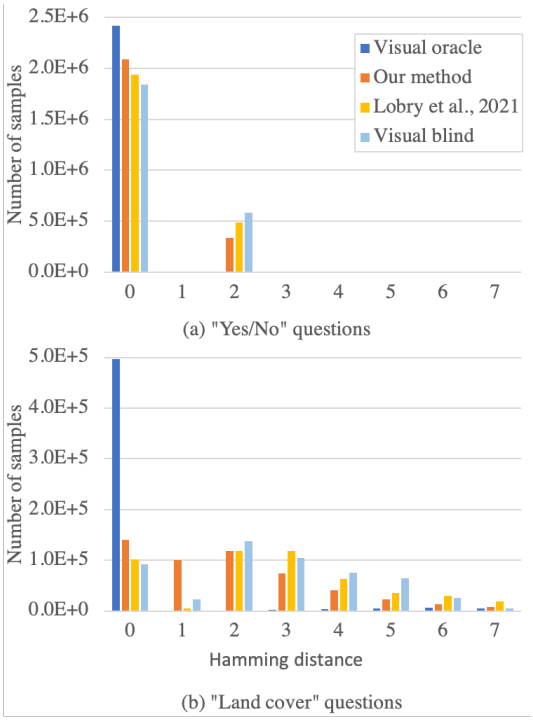

Figure 6. Distribution of per sample Hamming distances, separated by question type, for each method.

blind model. Our method outperforms [19] by achieving lower distances. While the average distance of the "Yes/No" questions fits the global tendency, the situation diverges slightly for the "Land cover" questions. Indeed, the HD for RSVQA [19] is higher (i.e. worse) than the *visual blind model*. Although not reflected in the global results, as the proportion of "Land cover" questions is much lower than the one of "Yes/No" questions (about 20% vs. 80% respectively), this result is troubling. It motivates the proposition of developing additional metrics to evaluate and challenge the performances of a model.

Figure 6 provides more details to the Hamming distance results by displaying their distributions per question type. The top (a) and bottom (b) fractions of the figure illustrate these distributions for the "Yes/No" and "Land cover" questions respectively. Again, the *visual oracle model* shows the best distribution with the large majority of samples at $HD = 0$. The *visual blind model* has the least number of samples with a perfect match compared to the other three cases. Our proposed method shows distributions more skewed towards 0 than [19] for both question types. Only two distances exist for the top (a) figure, indicating that the models most probably recognizes the question type and has a HD of 0 when predicting correctly, and 2 otherwise (binary swap for both "yes" and "no"). In the bottom (b) figure, a wider range of distances are seen and distributions are generally decreasing with higher distance value, to the exception of $HD = 1$. Higher distances do exist but in very small numbers and thus have been excluded from the figure for more clarity. Compared to [19], our method shifts its prediction towards smaller distances, especially $HD = 1$.

6. Conclusion

In this study, we proposed a prompted language model to address the remote sensing visual question answering task. The proposed method, *Prompt–RSVQA*, reframes the balance between the two modalities, vision and language, giving a leading role to the latter. The image is processed by a visual model, whose results are converted to text and used as visual context by the language model answering the question. Our results showed a 6% increase in accuracy compared to the baseline model RSVQA [19].

The vision and language models are fine-tuned separately before being assembled and run during inference. End-to-end training of the method is not trivial and is left to future research, as well as the exploration of architectures for the visual part (e.g. ResNet-152), or the usage of text generating modules enabling a wider set of answers.

In addition to improving the results on the RSVQA task, we demonstrated that this method allows to better evaluate the performances of each modality separately and to understand the influence of the visual performances on the RSVQA task. In particular, we have shown that a Transformer-based language model is sufficient for the RSVQA task. This is made possible by the semantic bottleneck, which is learned thanks to the BigEarthNet labels. In the development of future datasets, we thus encourage the storage of any additional information not directly contained in the image/question/answer triplets but that could still be extremely useful in later research.

Finally, we hope this first study will raise interest in the potential of using language as the reference modality in remote sensing VQA, instead of learning deep bi-modal embeddings. Context information originating from any type of data modality, given that it can be converted into a textual form, can be exploited in a prompting fashion, therefore allowing to leverage the latest pre-trained language models with no or light fine-tuning.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6077–6086, June 2018. ISSN: 2575-7075. 2

[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 IEEE international conference on computer vision, 2015. 1, 2

[3] Rishi Bommasani et al. On the Opportunities and Risks of Foundation Models. arXiv e-prints, 2021. 2

[4] Tom Brown et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901, 2020. 2

[5] 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 ECML-PKDD 2021 (MACLEAN workshop), 2021. arXiv: 2109.11848. 2

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, Computer Vision – ECCV 2020, Lecture Notes in Computer Science, pages 104–120, Cham, 2020. Springer International Publishing. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, Miami, FL, June 2009. IEEE. 3

[8] 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 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2, 3

[9] Rafael Felix, Boris Repasky, Samuel Hodge, Reza Zolfaghari, Ehsan Abbasnejad, and Jamie Sherrah. Cross-modal visual question answering for remote sensing data: The international conference on digital image computing: Techniques and applications (dicta 2021). In 2021 Digital Image Computing: Techniques and Applications (DICTA), pages 1–9, 2021. 2

[10] Lachezar Filchev, Lyubka Pashova, Vasil Kolev, and Stuart Frye. Challenges and solutions for utilizing earth observations in the" big data" era. arXiv preprint arXiv:2108.08886, 2021. 1

[11] Tianyu Gao, Adam Fisch, and Danqi Chen. Making Pre-trained Language Models Better Few-shot Learners. In Association for Computational Linguistics (ACL), 2021. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016. ISSN: 1063-6919. 3

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In NIPS 2014 Deep Learning Workshop, 2014. 3

[14] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A Good Prompt Is Worth Millions of Parameters? Low-resource Prompt-based Learning for Vision-Language Models. arXiv:2110.08484 [cs], Oct. 2021. arXiv: 2110.08484. 2

[15] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1988–1997, Honolulu, HI, July 2017. IEEE. 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 4

[17] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In Proceedings of the International Conference on Learning Representations (ICLR), 2020. 2

[18] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586, 2021. 2

[19] Sylvain Lobry, Begüm Demir, and Devis Tuia. RSVQA meets BigEarthNet: a new, large-scale, visual question answering dataset for remote sensing. In IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2021. 1, 2, 3, 4, 6, 7, 8

[20] 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, Dec. 2020. 1, 2

[21] Diego Marcos, Sylvain Lobry, Ruth Fong, Nicolas Courty, Remi Flamary, and Devis Tuia. Contextual semantic interpretability. In Asian Conference on Computer Vision (ACCV), Kyoto, Kapan, 2020. 4

[22] Adrien Moiret-Guigand and Gabriel Jaffrain. CLC 2018 and CLC change 2012-2018 validation report — copernicus land monitoring service. File, SIRS SAS, 2021. 3

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR, 18–24 Jul 2021. 2

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In NeurIPS, 2019. 3

[25] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 255–269, Online, Apr. 2021. Association for Computational Linguistics. 2

[26] Mike Schuster and Kaisuke Nakajima. Japanese and Korean voice search. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5149–5152, Kyoto, Japan, Mar. 2012. IEEE. 5

[27] Shaden Smith et al. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. arXiv:2201.11990 [cs], Feb. 2022. arXiv: 2201.11990. 2

[28] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis, 18:1–25, 2010. 5

[29] Gencer Sumbul, Marcela Charfuelan, Begum Demir, and Volker Markl. Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, pages 5901–5904, Yokohama, Japan, July 2019. IEEE. 3, 6

[30] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-Structured Representations for Visual Question Answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2017. 2

[31] Romal Thoppilan et al. LaMDA: Language Models for Dialog Applications. arXiv:2201.08239 [cs], Feb. 2022. arXiv: 2201.08239. 2

[32] Devis Tuia, Ribana Roscher, Jan Dirk Wegner, Nathan Jacobs, Xiao Xiang Zhu, and Gustau Camps-Valls. Towards a Collective Agenda on AI for Earth Science Data Analysis. IEEE Geoscience and Remote Sensing Magazine, 9(2):88–104, June 2021. arXiv: 2104.05107. 1

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. 3

[34] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium, 2018. Association for Computational Linguistics. 3

[35] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. UFO: A UniFied TransfOrmer for Vision-Language Representation Learning. arXiv:2111.10023 [cs], Nov. 2021. arXiv: 2111.10023. 2

[36] Yonghui Wu et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144 [cs], Oct. 2016. arXiv: 1609.08144. 5

[37] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. arXiv:2109.05014 [cs], Sept. 2021. arXiv: 2109.05014. 2

[38] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked Attention Networks for Image Question Answering. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 21–29, Las Vegas, NV, USA, June 2016. IEEE. 2

[39] Lu Yuan et al. Florence: A New Foundation Model for Computer Vision. arXiv:2111.11432 [cs], Nov. 2021. arXiv: 2111.11432. 2

[40] Zhenghang Yuan, Lichao Mou, Zhitong Xiong, and Xiaoxiang Zhu. Change detection meets visual question answering, 2021. 2

[41] Zhenghang Yuan, Lichao Mou, and Xiao Xiang Zhu. Self-paced curriculum learning for visual question answering on remote sensing data. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pages 2999–3002, 2021. 2

[42] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal Neural Script Knowledge Models. Technical report, June 2021. Publication Title: arXiv e-prints ADS Bibcode: 2021arXiv210602636Z Type: article. 2

[43] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering, 26(8):1819–1837, 2014. 5

[44] 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, pages 1–14, 2021. 2

[45] Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. Adapting Language Models for Zero-shot Learning by Meta-tuning on Dataset and Prompt Collections. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2856–2878, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2

[46] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 19–27, Santiago, Chile, Dec. 2015. IEEE. 3