Environmental Computational Science
and Earth Observation Laboratory

Master thesis subject

OK-RSVQA: Towards open knowledge remote sensing visual question answering

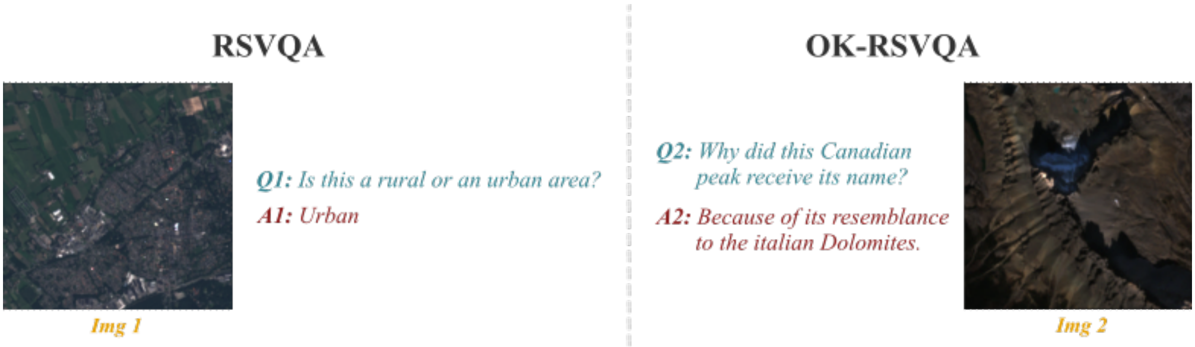
RSVQA
OK-RSVQA
Q1: Is this a rural or an urban area?
A1: Urban
Img 1
Q2: Why did this Canadian peak receive its name?
A2: Because of its resemblance to the italian Dolomites.
Img 2

Figure: Examples of (Image, Question, Answer) triplets in RSVQA. Left: Sample from RSVQA dataset (Lobry et al. 2020). Right: The answer to the question cannot be found given the image alone, the image and the question serve as context to find the answer in an external knowledge source. This setting is called Outside Knowledge RSVQA (OK-RSVQA).

Context

Visual question answering (VQA) is the task of answering questions in natural language about the visual content of an image. Inherently, a system that is trained to perform VQA needs to be able to understand the question, analyse the content of the image and provide a coherent answer in natural language. VQA has a wide range of applications in the real world, for example, by allowing visually impaired people to make inquiries about their surroundings. In remote sensing (RS) applications, VQA is of particular interest. It would enable non-expert users to extract useful information from RS imagery, providing the possibility to a larger public to use these powerful resources.

In the last years, several works have explored VQA in the RS field (RSVQA), proposing new benchmarks and methods and showing the potential of VQA for remote sensing. However, current approaches and datasets for RSVQA are still limited to the objects and features that can be directly extracted from the image (e.g., presence of objects or landcover types, counting objects); and they do not allow us answering questions that need more general knowledge or reasoning beyond what is in the image. Fig. 1 shows examples of possible questions to be asked in the context of RSVQA. On the left, a system only needs to understand the content of the image to provide an answer to the question. On the other hand, to answer the question on the right, a system would need an external source of knowledge and visual content to answer the question correctly.

This problem is known in the literature as Outside Knowledge VQA (OKVQA), where the aim is to build a VQA system that is able to rely on external knowledge sources to answer image-related questions.

Therefore, this project aims to create the first large-scale VQA dataset in remote sensing, containing both outside knowledge and image-grounded questions, and provide some baselines to tackle this task.

Objectives

- Studying and understanding the main challenges of VQA, OKVQA and RSVQA
- Using existing resources to create questions related to remote sensing images and external knowledge
- Providing baselines for OK-VQA in remote sensing

Requirements

- Python programming skills
- Machine learning/deep learning experience (in particular, vision and/or NLP).
- Familiarity with GIS systems is a plus
- Willingness to learn and ability to work independently

Literature

- S. Lobry, D. Marcos, J. Murray, and D. Tuia. *“RSVQA: Visual Question Answering for Remote Sensing Data”*. IEEE Transactions on Geoscience and Remote Sensing. 2020.
- C. Chappuis, V. Zermatten, S. Lobry, B. Le Saux, D. Tuia. *“Prompt-RSVQA: Prompting visual context to a language model for Remote Sensing Visual Quesiton Answering”.* IEEE Conference on Computer Vision and Pattern Recognition – EarthVision workshop. 2022.
- K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. *“OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge”.* IEEE Conference on Computer Vision and Pattern Recognition. 2019.

Contact

Prof. Devis Tuia, devis.tuia@epfl.ch