Text and remote sensing images interactions

Master internship

General information

- Keywords: : Computer vision, Natural Language Processing, Deep Learning, Remote sensing, Multi-modality
- Duration of the internship: 6 months (standard stipend). Start: from March 2022.
- Institute: Université de Paris, Laboratoire d'Informatique Paris Descartes (LIPADE), équipe Systèmes Intelligents de Perception
- Location: 45 rue des Saints-Pères, 75006 Paris
- Supervisors: Sylvain Lobry, Camille Kurtz, Laurent Wendling
- Application: please send an email to sylvain.lobry "at" u-paris "dot" fr with the subject "[Internship M2 2022_1] FirstName LastName" containing:
  - an updated CV;
  - your grades and ranking of your master degree;
  - a cover letter;
  - the contact information of a teacher/supervisor willing to write a recommendation letter.

  The position is opened until filled.

Motivation

In recent years, remote sensing images have become more available than ever thanks to important efforts coming from the public and private sectors. For instance, the European Union's Copernicus program provides free access to Synthetic Aperture Radar (SAR) and multi-spectral data. In addition to governmental initiatives, companies (e.g. Planet Labs) also provide very-high resolution images on a global scale on a daily basis. Remote sensing images contain information which is already used, among others, to track climate change, improve security and to understand and manage the environment. Exploiting the different levels of information provided by the wide range of remote sensing modalities is an active field of research and used in many remote sensing applications [1]. However, the interpretation of remote sensing data is generally made by experts and often involves manual processing. With the increasing amount of data, the manual interpretation becomes a limiting factor impacting the delay at which information is extracted, but also the domains in which such data can be used. For specific applications, the remote sensing community has been developing ad hoc automatic methods. As such, these works can only address either general applications (e.g. pollution monitoring) or ones with direct financial interest. We argue that the information contained in remote sensing images can be of interest to a much larger public: journalists could retrieve such data to understand, follow and report on wars and the effects of climate change or local governments could use this data in their decision process and studies. While the data is here, the general audience do not always have either the technical knowledge to extract the information of interest or the capacity to fund research to do so. Enabling information extraction from remote sensing data through a non-technical and common interface would be a way to allow the general audience to directly benefit from this data. In this project, we propose to explore the use of natural language as an interface.

Background

Interactions between textual features and images is a rising topic in the machine learning and computer vision communities. In particular, these interactions are essential components of tasks such as image captioning (IC) [2], image querying (IQ) [3] or Visual Question Answering (VQA) [4]. These tasks are particularly relevant when used with remote sensing data. Indeed, image querying has been a task of interest in the remote sensing community since the creation of massive remote sensing images databases as a way to explore them through natural language [5]. On the other hand, VQA has only been recently proposed in the remote sensing community [6] and has been identified as one of 6 potential game-changers in the field of Artificial Intelligence for Earth Science [7]. It aims at answering in English to a question (in English as well) about a remote sensing image. Since the introduction of this task in the remote sensing community in 2019, numerous research works have explored the construction of large databases for the training of supervised models [6, 8, 9] as well as the models themselves [6, 10, 11, 12]. However, no method has been proposed to jointly learn a common representation for text and remote sensing images, regardless of the final task. In this internship, we propose to work towards the creation of such a method.

Objectives

The work to be conducted during the proposed M2 internship will lead to the following three contributions, combining generic, open-sourced contributions (contribution A) and exploratory works (contributions B and C):

- Contribution A: Participation in the construction of a large database using automatic methods. By deriving available annotations (e.g. from OpenStreetMap or IGN's BDTOPO) to a common representation, we will be able to study in a common framework the tasks of IC, IQ and VQA. Due to the large coverage needed, problematics of big data will be studied.
- Contribution B: Enabling links between different levels of information in the ground truth. On a small subset of images, we will propose a dense and structured annotation inspired by Visual Genome [13]. We will study how this information can be translated to the common representation extracted in contribution A, or task specific ground truth.
- Contribution C: Understanding the dataset. With this new dataset, we will train one baseline model for one of the tasks of interest (IC, IQ, VQA) to understand what knowledge can be extracted. Additionally, an analysis on the limitations of the base model will be conducted.

Together, the 3 contributions will allow to introduce a new, large-scale dataset to the remote sensing community. As such, the resulting dataset will be made openly and freely available. Furthermore, this work will lay the foundations for future methodological research on the tasks of IC, IQ and VQA. There will be the opportunity to further work on these topics in a PhD funded by the French research agency (ANR).

Desired background for the candidate

We are looking for a student in final year of MSc or engineering school. The ideal candidate would have a theoretical background in computer vision and be proeficient in programming with Python. Knowledge in geographic information sciences, natural language processing and an interest in handling large amount of geo-coded data is a plus.

Bibliography

[1] Mauro Dalla Mura et al. "Challenges and opportunities of multimodality and data fusion in remote sensing". In: *Proceedings of the IEEE* 103.9 (2015), pp. 1585–1601.

[2] Quanzeng You et al. "Image captioning with semantic attention". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4651–4659.

[3] Huafeng Wang et al. "Deep learning for image retrieval: What works and what doesn't". In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE. 2015, pp. 1576–1583.

[4] Stanislaw Antol et al. "Vqa: Visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

[5] Klaus Seidel et al. "Query by image content from remote sensing archives". In: *IGARSS'98. Sensing and Managing the Environment. 1998 IEEE International Geoscience and Remote Sensing. Symposium Proceedings.(Cat. No. 98CH36174)*. Vol. 1. IEEE. 1998, pp. 393–396.

[6] Sylvain Lobry et al. "RSVQA: Visual question answering for remote sensing data". In: *IEEE Transactions on Geoscience and Remote Sensing* 58.12 (2020), pp. 8555–8566.

[7] Devis Tuia et al. "Toward a Collective Agenda on AI for Earth Science Data Analysis". In: *IEEE Geoscience and Remote Sensing Magazine* 9.2 (2021), pp. 88–104.

[8] Sylvain Lobry, Begüm Demir, and Devis Tuia. "RSVQA Meets Bigearthnet: A New, Large-Scale, Visual Question Answering Dataset for Remote Sensing". In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE. 2021, pp. 1218–1221.

[9] Maryam Rahnemoonfar et al. "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding". In: *IEEE Access* 9 (2021), pp. 89644–89654.

[10] Christel Chappuis et al. "How to find a good image-text embedding for remote sensing visual question answering?" In: *MACLEAN Workshop at ECML*. 2021.

[11] Xiangtao Zheng et al. "Mutual Attention Inception Network for Remote Sensing Visual Question Answering". In: *IEEE Transactions on Geoscience and Remote Sensing* (2021).

[12] Sylvain Lobry et al. "Better Generic Objects Counting When Asking Questions to Images: A Multitask Approach for Remote Sensing Visual Question Answering". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 2020, pp. 1021–1027.

[13] Ranjay Krishna et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations". In: *International journal of computer vision* 123.1 (2017), pp. 32–73.