Partial latent encoding of multi-spectral data

Master internship

General information

- Duration: 6 months (standard stipend). To start between February and April 2024.
- Institutes: Université Paris Cité, Laboratoire d'Informatique Paris Descartes (LIPADE), team Systèmes Intelligents de Perception and team EVERGREEN (Inria, INRAE, Cirad)
- Location: 45 rue des Saints-Pères, 75006, **Paris** (LIPADE)
  or 500, rue Jean François Breton, 34090, **Montpellier** (EVERGREEN)
- Supervision: Sylvain Lobry, Camille Kurtz, Laurent Wendling (LIPADE), Diego Marcos (Inria), Dino Ienco (INRAE)
- **Application: Please apply on https://recrutement.inria.fr/public/classic/fr/offres/2023-06969. The position is open until filled, and full consideration will be given to application received before 15/12/2023.**

Proposed topic

Context

By using location on the Earth's surface as the common link between different modalities, a geo-spatial foundation model would be able to incorporate a variety of data sources, including remote sensing imagery, textual descriptions of places, and features in maps. Leveraging the large amounts of available unlabeled geo-spatial data from these different sources, the GEO-ReSeT [1] (Generalized Earth Observation with Remote Sensing and Text) ANR project has the objective to learn a better representation of any geo-spatial location and convey a semantic representation of the information. Such a foundation model has the potential to revolutionize Earth observation by allowing for few or zero-shot solutions to classical problems such as land-cover and land-use mapping, target detection, and visual question answering. It will also be useful for a wide range of applications with a geo-spatial component, including environmental monitoring, urban planning and agriculture. By leveraging several data modalities, this foundation model could provide a more comprehensive and accurate understanding of the Earth's surface, enabling more informed decisions and actions. This will be particularly valuable for new potential users in sectors such as journalism, social sciences or environmental monitoring, who may not have the resources or expertise to collect their own training datasets and develop their own methods, thus moving beyond open Earth observation data and democratizing the access to Earth observation information.

Work to be done

The work to be conducted during the proposed M2 internship will contribute to the ambition of the GEO-ReSeT ANR project by proposing a new methodology for projecting multi-modal data of different natures to a common latent space. One classical way to achieve this is through a contrastive self-supervised learning approach. A feature extractor for each modality is trained through a contrastive loss. This loss ensures that similar examples (in the case of geo-spatial data, from the same geographical location) are close in the feature space, while dissimilar examples are projected far away [1]. These self-supervised models can then be used on downstream tasks through linear probing.

This approach tends to work well on natural images and has been successfully on geo-spatial data, such as remote sensing image [2]. However, retaining the particularities of each modality, each given partial informationof the underlining reality, is a challenge [2]. A first solution to this issue is proposed in [3]. In this work, the authors propose to learn factorized representations of each modalities.

In this work, our objective is to to explicitly model which part of the latent space is concerned with each of the modalities as shown in Figure 1. We propose to achieve this objective by modeling the uncertainty on the feature representation of each modality. The work to be performed in this internship will lead to the following three contributions:

[1] https://geo-reset.sylvainlobry.com/

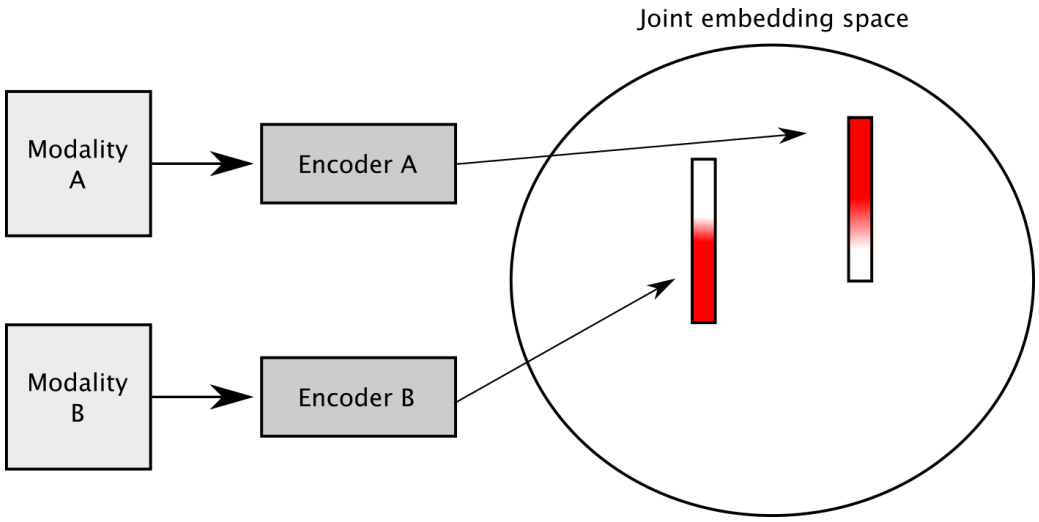

Figure 1: Illustration of the proposed methodology. The objective is to predict a representation for each modality in a joint embedding space with a characterization of the certainty of the representation (in red) or uncertainty (in white).

- Contribution A: the candidate will propose a methodology for self-supervised learning of a joint embedding of multi-modal data. This methodology will explicitly encode the uncertainty during the learning process and at inference.
- Contribution B: the candidate will choose and implement a baseline that can be used for comparison with the proposed method.
- Contribution C: the candidate will choose a downstream task on which the proposed methodology can be evaluated. A detailed evaluation will be conducted.

Desired background

We are looking for a Master 2 student or final year of MSc, or engineering school in computer science. The ideal candidate should have knowledge in image processing, computer vision, natural language processing, geo-information sciences, Python programming and an interest in handling large amount of data, in particular remote sensing.

Bibliography

[1] Heechul Jung et al. "Contrastive self-supervised learning with smoothed representation for remote sensing". In: *IEEE Geoscience and Remote Sensing Letters* 19 (2021), pp. 1–5.

[2] Zhixi Feng et al. "Cross-modal contrastive learning for remote sensing image classification". In: *IEEE Transactions on Geoscience and Remote Sensing* (2023).

[3] Paul Pu Liang et al. "Factorized Contrastive Learning: Going Beyond Multi-view Redundancy". In: *arXiv preprint arXiv:2306.05268* (2023).