Cross-modal reciprocal learning: location as supervision

Laboratories: LASTIG lab. (Univ Gustave Eiffel, IGN, ENSG) & IMAGINE (ENPC)

Localisation: LASTIG, IGN-ENSG, Saint-Mandé, France

Supervision: Loic Landrieu, PhD HDR; Clément Mallet, PhD HDR; Nicolas Gonthier, PhD

Starting Date: September 2023

Keywords: Deep Learning, Self-supervision, Large-Scale, Land Cover, Geospatial Imagery, Open-Source, Environmental Monitoring

Development Environment: Linux, Python, PyTorch

1 Context

Earth Observations (EO) nowadays refer to sensors with different modalities capturing rich and complementary information. EO are typically large-scale, multimodal, and multi-resolution. They often display a complex structure in which the spatial, temporal, and spectral dimensions are entangled [4]. This structure is influenced by phenomenons operating at widely different scales, complex domain shifts (the distribution and appearance of certain classes vary significantly across regions and throughout the year), atmospheric conditions, and cumulative weather, resulting in inherently non-stationary processes. While these characteristics complexify the analysis of EO, their absolute spatial and temporal referencing of remote sensing allows us to align different acquisitions easily [3].
The deep learning paradigm is gradually being adopted as an indispensable tool for the automatic analysis of EO. However, the state-of-the-art solutions suffer from multiple limitations: (1) existing deep-based models are not well-suited to the structure of EO, (2) large-scale annotation is unrealistic, (3) the limited spatio-temporal extent and task-specificity of existing datasets and models result in poor generalization, (4) most annotated datasets remain focused on specific areas and environments, biasing the community towards particular architectures.
We propose to address the previous limitations by developing a Foundation Model of Earth Observations, trained on extensive unannotated data. Such a model should be rich, expressive, compact and efficient for most domain shifts and geographical bias. We will explore the idea of reciprocal cross-modal training to leverage the multimodality and georeferential alignment of EO data.

This PhD aims to take advantage of the amount of spatially consistent Earth Observation data to **build a Foundation Model on which downstream tasks can be easily applied** such as land-cover classification or biogeophysical variable estimation.

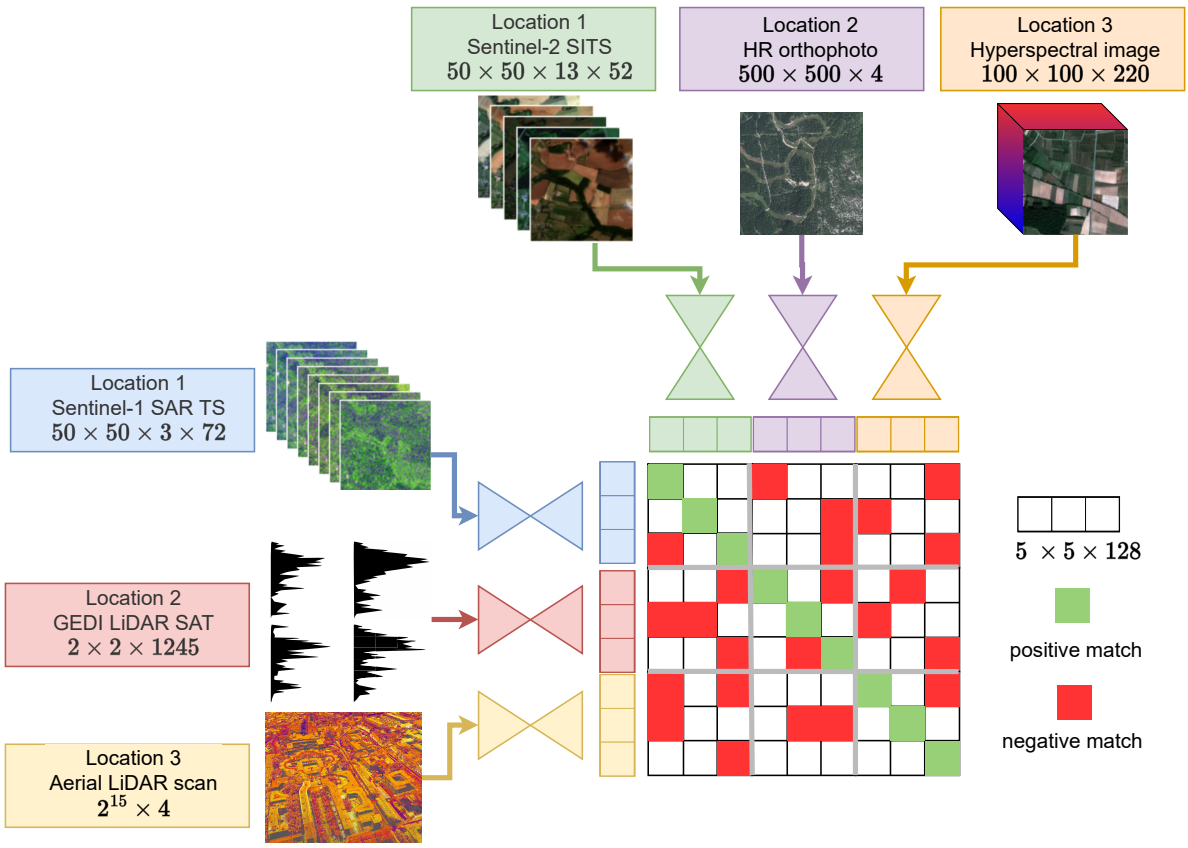

Figure 1: **Reciprocal cross-modal learning.** Cooperation between multiple heterogeneous EO data sources using a common spatial support and semantic information as supervision.

2 Objective

Foundation models refer to large neural networks trained on a vast amount of data and widely used for further applications or research [2]. Such models can be built upon without having to be trained from scratch. This strategy both saves computation time and provides expressive features/state-of-the-art encoders, and decreases the annotation and hardware requirements usually associated with large modern networks. We propose a new learning paradigm dubbed "*Cross-modal reciprocal learning*" to train such foundation models from *unannotated multi-modal EO*.

Text-Image contrastive pretraining has shown impressive results in computer vision [7, 5], leading to expressive and influential foundation models. Contrastive learning has been recently explored for EO by exploiting spatial alignment across time series [1] or for cross-modal localization [6]. We propose generalizing text-image contrastive learning to the multi-modal setting with spatially aligned observations. We require the features extracted from acquisitions of an area through different modalities to be more similar than any descriptors of another area. By forcing spatial alignment across sensors, the features must describe the only shared latent variable: the actual semantics of the acquired area (*e.g,* road, building, plant species).

Generalizing contrastive learning to the multi-domain setting raises several theoretical and technical challenges. First, the classic two-modalities formulation leads to an exponential complexity w.r.t the number of sensors, quickly becoming impractical and requiring us to develop an adapted loss and sampling procedure. Second, cross-modal learning implies the simultaneous training of several large networks and the manipulation of costly multi-modal batches. This raises technical issues such as inefficiency in memory use and prolonged training times. Lastly, if we only reward the encoders for spatial alignment across modalities, they may be discouraged from retaining sensor-specific informa-

tion. This would result in weaker individual representations discarding the strength of each sensor at the benefit of the lowest common denominator. This critical concern may be mitigated with multi-task learning and self-supervision.

Depending on the application cases, we will focus on several modalities among the six following ones: Sentinel-2 Optical Time Series, Sentinel-1 Radar Time Series, High Resolution orthoimages, hyperspectral images, aerial LiDAR, and spaceborne LiDAR.

The PhD project also includes application cases related to the exploitation of the designed model. We will focus on standard downstream tasks such as multi-domain land-cover classification or crop mapping, and biogeophysical variable estimation (*e.g.*, biomass estimation).

L'approche proposée pourra servir de base à la construction de nombreux modèles spécialisés et, potentiellement, à une utilisation intensive des données issues du Lidar HD. Avis très positif de l'équipe Strudel.

Profile

- Master 2 student in computer science, applied mathematics, or remote sensing.
- Familiarity with computer vision, machine learning, and deep learning.
- Mastery of Python, familiarity with PyTorch;
- Curiosity, rigour, motivation;
- (Optional) Familiarity with self/weak/un/contrastive learning;
- (Optional) Experienced with aerial/satellite sensor technology and land cover/land use prediction models.

Contact

Send a CV and a short letter of purpose ($\sim$20 lines max) stating your interest in this PhD project and the relevance of your experience to `loic.landrieu@ign.fr` and `clement.mallet@ign.fr`.

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. *ICCV*, 2021.

[2] Rishi Bommasani et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021.

[3] Danfeng Hong, Naoto Yokoya, Gui-Song Xia, Jocelyn Chanussot, and Xiao Xiang Zhu. X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:12–23, 2020.

[4] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112:102926, 2022.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[6] W.-H. Tseng, H.-A. Lê, A. Boulch, S. Lefèvre, and D. Tiede. Croco: Cross-modal contrastive learning for localization of earth observation data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2022:415–421, 2022.

[7] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.