Earth observation at scale is a complex task. Satellite data and generic geolocated imagery provide a very detailed view of the surface of the Earth, and the temporal dimension of image acquisition allows to study Earth surface processes with unprecedented detail. Usually, the practitioner collects ground truth data about land-cover classes or processes of interest, selects appropriate imagery, and trains a classifier on such pairing so that the same land-cover classes can be inferred from previously unseen images, covering new geographical areas or acquired at different time instants.

Although this standard procedure is satisfactory for localized and small scale studies, it becomes suboptimal when moving at the scale of a country or a continent, or generally when the amount of data to be ingested and processed is just too large and diverse. Obvious issues are the lack of ground truth annotations, a consequent shift of the class-conditional spectral distributions (e.g. a forest in Switzerland has different spectral responses than forests in Brasil, due to different vegetation and atmospheric properties).

Recent research efforts, sustained by the ability of deep learning models to provide good generalization, aim at optimally transfer pre-trained models on (spatio-temporally) disjoint data, by either fine tuning with little labels or by few-shot learning. However, those approaches require that available ground truth annotations is not scarce and representative of the target distribution, which in many practical scenarios is unrealistic. Potential ways to address this
problem is to learn optimal data representations or to learn models which are trained on proxy

tasks which are similar to the final, downstream task. Since the data domain only shares few

similarities with natural images, the inductive bias in the models needs to be favouring specific

setting that are uncommon to standard computer vision tasks on common benchmarks, e.g.

image recognition on ImageNet [1], instance segmentation on MS COCO lin14eecv or semantic

segmentation on PASCAL VOC [2].

In this project, we aim at developing and potentially improving upon recent approaches

aiming and global, pretrained, spatio-temporal representation for satellite and overhead images.

To this end, we will explore coupling of unsupervised embedding methods, specifically

Tile2Vec [3], with few-shot learning [4] and meta-learning [5], where we aim at using this

base pre-trained representation to learn and encode in such model a set of tasks related to a

downstream application.

Application 1 not only focuses on abovehead imagery, but also include into the pipeline

multiple modalities, and specifically street-level imagery [6]. The application focus will be on

urban areas, where street level images are abundant and representative of aspects not visible

from above.

Application 2 aims at refining this framework for potential Earth observation applications

at scale [7], such as crop monitoring or land-cover classification [8].

This project is ambitious, but it has to first be separated into one of the two applications,

which then can further be separated into distinct tasks and each one of those explored independ-

ently (e.g. learning the representation, meta-learning, few-shot learning). Any contribution

in each one of those aspects is worthy, and potentially of interest for the whole geospatial and

Earth observation community.

Additional information

• Difficulty of the project: Challenging (but you won’t be alone)

• What will you learn? Lots of deep learning flavours (based on CNNs), good scientific

re...


