Heterogeneous Data and Large Representation Models in Science

Contribution ID: **39**   Type: **Oral presentation**

Keynote Address: Multimodal Pretraining for Astrophysical Foundation Models

Wednesday 2 October 2024 10:55 (45 minutes)

Deep Learning has seen a recent shift in paradigm, from training specialized models on dedicated datasets, so-called Foundation Models, trained in a self-supervised manner on vast amounts of data and then adapted to solve specific tasks with state-of-the-art performance. This new paradigm has been exceptionally successful not only for large language models (LLMs) but in other domains such as vision models. However applications of this new approach in astrophysics are still very scarce, for reasons ranging from new architectures to the (surprising) lack of availability of suitable large scale datasets.

In this talk, I will discuss our recent work on deploying such a Foundation Model approach in the context of representation learning for astronomical photometric and spectroscopic observations of galaxies. Our aim is to embed these inhomogeneous observations (e.g. different types of measurements, different instruments, etc...) into a shared embedding space, in a completely self-supervised manner. These embeddings can then be used for a variety of downstream applications (e.g. redshift estimation, morphology classification, estimating physical properties) with very simple machine learning methods and reach near optimal performance.

More specifically, I will present our AstroCLIP method which allows us to align embeddings between data modalities, but also our more recent and ongoing work on building early-fusion multimodal models relying on modality-specific tokenizers and a joint large transformer model.

Contribution length

Long

Primary author: LANUSSE, François (CNRS, UMR AIM / Flatiron Institute)

Presenter: LANUSSE, François (CNRS, UMR AIM / Flatiron Institute)