Foundation Models for Astronomy

Yihan Tao

China-VO

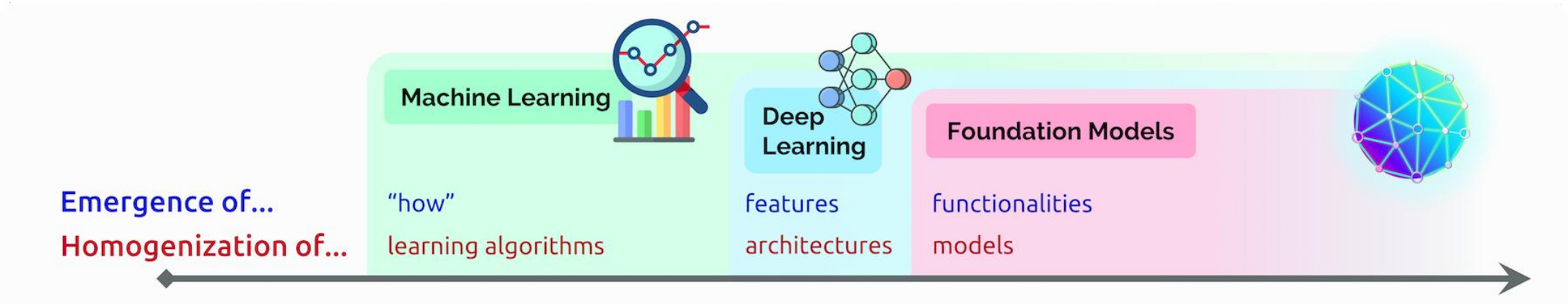
KDIG Session, IVOA Interop Meeting, Bologna, Italy
9 May 2023

Outline

- Concept of foundation models
- Related work in Astronomy
- Potential of foundation models for Astronomy

What is “Foundation Models”

- Foundation models are models trained on large and broad data, generally through self-supervised learning and can be adapted to a wide range of downstream tasks.
- Based on conventional deep learning and transfer learning

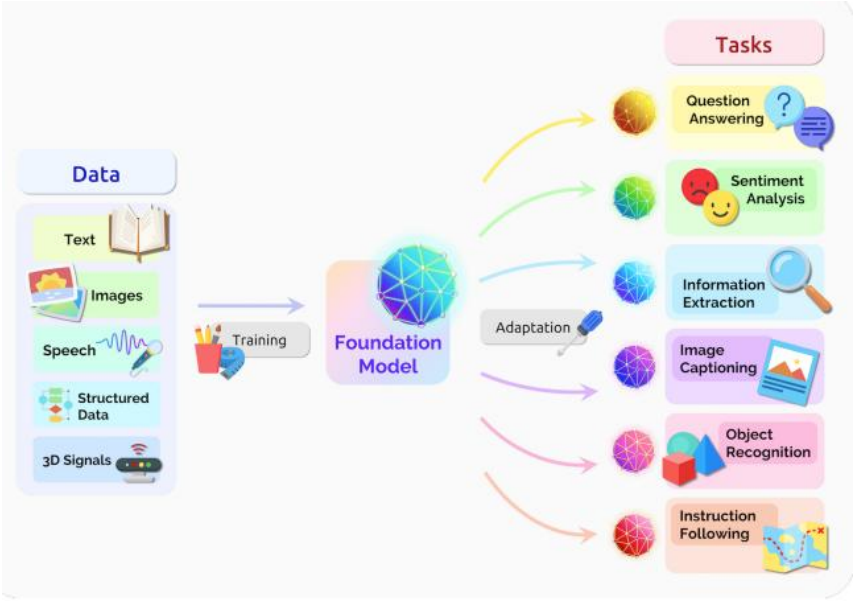


- Related concepts:
  - Large language models (LLMs) - focus on NLP tasks
  - Self-supervised model – describe the training approach
  - Pre-trained model – describe model building strategy

Rishi Bommasani, Percy Liang et al. On the Opportunities and Risks of Foundation Models. Stanford Center for Research on Foundation Models (CRFM)

Why Foundation Models

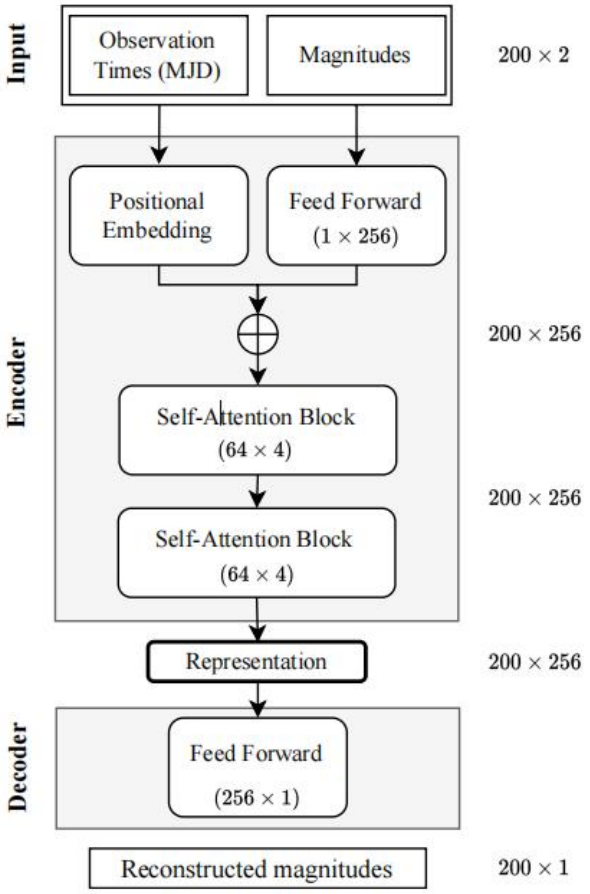
- Powerful basis for AI services and applications
  - ChatGPT = foundation model (GPT-4) + prompt tuning + reinforcement learning
- Homogeneous representation for various tasks
- Using large unlabelled data



Rishi Bommasani, Percy Liang et al. On the Opportunities and Risks of Foundation Models.

Related work in Astronomy

- ASTROMER: A transformer-based embedding for the representation of light curves
  - pre-trained on millions of light curves from different surveys (MACHO, OGLE, ATLAS)
  - representation to create informative light curves embeddings
  - finetuned for solving downstream tasks, e.g. classification of variable stars, predicting physical parameters



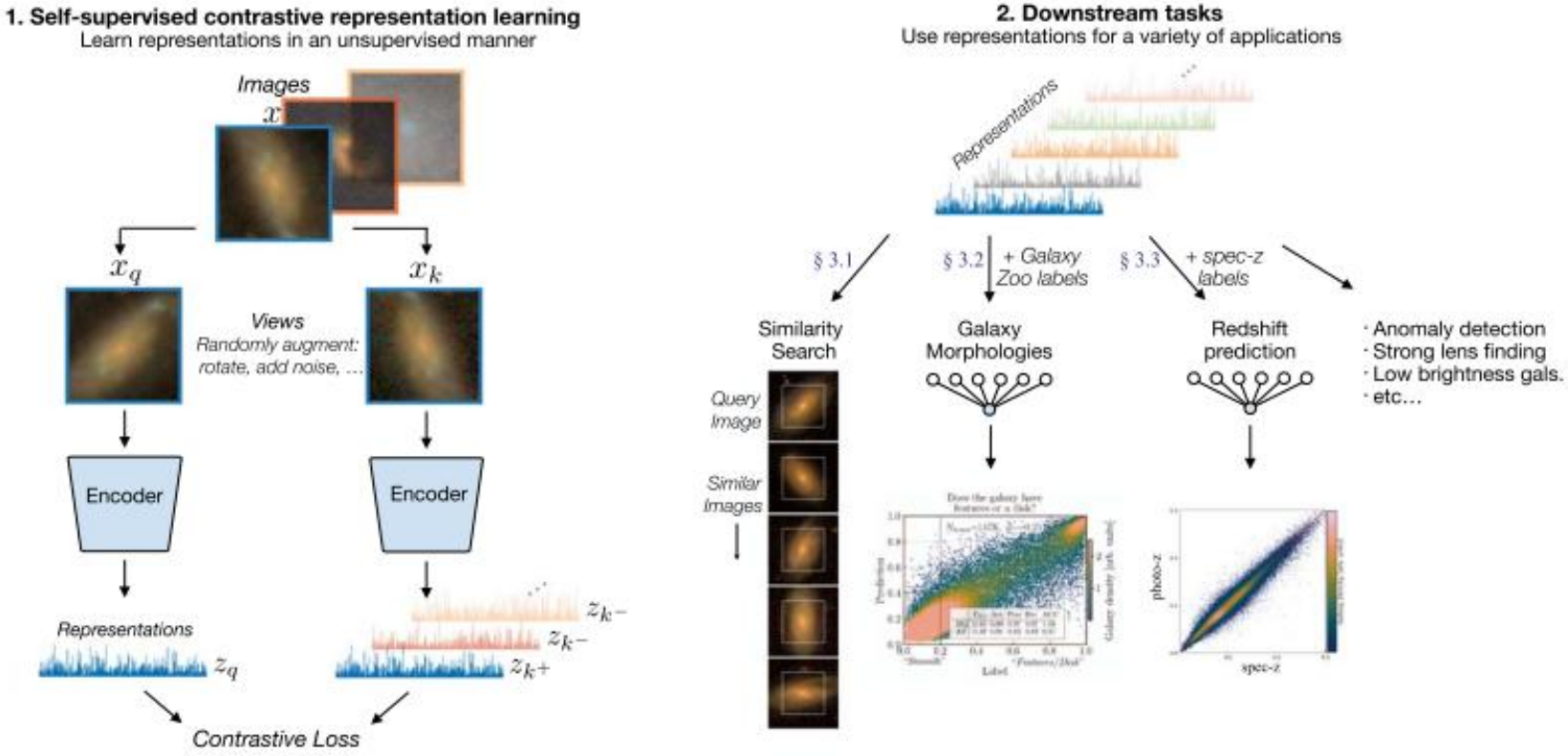
ASTROMER

https://www.stellardnn.org/projects/astromer/index.html

C. Donoso-Oliva et al. ASTROMER: A transformer-based embedding for the representation of light curves.

Related work in Astronomy

- Self-supervised Representation Learning for Astronomical Images
  - multiband galaxy photometry from the Sloan Digital Sky Survey (SDSS) to learn image representations



Md Abul Hayat et al. Self-supervised Representation Learning for Astronomical Images.

Related work in Astronomy

- Towards Galaxy Foundation Models with Hybrid Contrastive Learning
  - 552k labelled and 1.34m unlabelled galaxies from five telescopes and four Galaxy Zoo campaigns
  - ML-friendly galaxy datasets for major Galaxy Zoo (https://github.com/mwalmsley/galaxy-datasets)

Name	Method	PyTorch Dataset	Published	Downloadable	Galaxies
Galaxy Zoo 2	gz2	GZ2	☑	☑	~210k (main sample)
GZ Hubble*	gz_hubble	GZHubble	☑	☑	~106k (main sample)
GZ CANDELS	gz_candels	GZCandels	☑	☑	~50k
GZ DECaLS GZD-5	gz_decals_5	GZDecals5	☑	☑	~230k (GZD-5 only)
GZ Rings	gz_rings	GZRings	☒	☑	~93k
GZ DESI	gz_desi	GZDesi	☒	WIP	WIP
CFHT Tidal*	tidal	Tidal	☑	☑	1760 (expert)

Mike Walmsley et al. Towards Galaxy Foundation Models with Hybrid Contrastive Learning

Potential of foundation models for Astronomy

- Foundation models mark the beginning of a new era in machine learning and artificial intelligence.
- How can it serve Astronomy
  - Apply LLMs to Astronomy papers, integrate domain knowledge for searching and question answering tasks
  - General representation for different data types (images, spectra, time series, catalogue etc.) for astronomical data analysis tasks
  - To serve as research assistant for more complicated task, e.g. plan for observation, generate report and figures. automatically decompose the task (AutoGPT)
- From VO perspective
  - Make the data ready - Standardize the representation for astronomical data of different type and from different sources?