Deep Generative Model for Stellar Spectra in the Wild

TL;DR: Build a data-driven deep learning model capable of generating a stellar spectrum.

Motivation

Our Milky Way and nearby satellite galaxies contain billions of stars which can be observed by our telescopes. An excellent fingerprint of a star is its spectrum, i.e. the intensity in function of wavelength, representing complex physical processes of underlying stellar properties. Having millions of stellar spectra help us determine the chemical evolution of our Milky Way, its historical dynamical evolution and trace the history of the elements even on earth.

Today multiple observatories are equipped with dedicated instruments capable of collecting thousands of spectra for each exposure. This makes the traditional manual analysis of stellar spectra basically impossible at the scale of millions of stellar spectra that are now publicly available.

Several research groups have put together dedicated software for stellar spectra analyses, for each telescope. Some of them are based on data-driven approaches such as convolutional neural networks. However these dedicated spectra analyses often come with their own biases, such that a given star observed by two telescopes and analyzed by two different groups will show different derived physical properties such as temperatures, gravity, or elemental abundances. Moreover, they are limited in range of wavelengths, and resolution by either the telescope or the synthetic modeling.

Because of the complexities in the physical interior and stellar atmosphere, building a realistic physical simulation that can reproduce all types of stellar spectra for any type of star, at any wavelength from the UV and the infrared has not been done yet. Maybe a machine could learn a model, at the cost of some of its physical (human) interpretability.

Project

The goal of this project will be to implement an unsupervised deep generative model, from the diversity of spectra at native resolution and wavelengths from at least 2 different telescopes. The resulting model should be able to generate a realistic stellar spectrum quickly.

The project will allow students to learn about generative modeling, signal processing and statistical techniques and possibly produce innovative research. This is an ambitious project but with good organization, technical skills and communication, can be done.

Plan suggestion:

1. Get vaguely familiar with what a stellar spectrum data: understand the format,

how to plot the spectrum. Gather spectra from various resolution and wavelength ranges. Possible data: the GALAH Survey (DR3), the APOGEE Survey (DR17), the Gaia-ESO Survey (DR4).

2. Get familiar with irregular spacing data. How can one deal with a spectrum at different wavelength coverage and resolution? Study how one could possibly do it. Read about position encodings. Could we use wavelength encodings? How would it help?
3. A possible generative model for irregular spaced data are the ones based on coordinate-based generative models. Read about Gradient Origin Networks, Could we use this model? What are its advantages vs. a VAE, a GAN?
4. Implement a generative model using the spectra. Select the proper architect (convolutional and/or transformer?).
5. Add simple physics for better interpretability. Consider a generative model that produces a spectrum in rest frame, and apply radial velocity to it before comparing to a real spectrum, i.e. wavelength_observed = (1+velocity/c)*wavelength_restframe.
6. Analyze its latent space and how their code influences parts of the spectrum. Does it relate to specific chemical abundance features?

These are only suggestive steps. A mid-term project could stop at 4.

Computing resources (~2-5TB + GPU) can be made available.

Please consult for more details!

Some references:

1. GALAH Survey: https://www.galah-survey.org/
2. APOGEE Survey: https://www.sdss.org/dr17/irspec/
3. Gradient Origin Networks: https://cwkx.github.io/data/GON/ - Basic implementation in pytorch: