Physics Informed Machine Learning in Astrophysics

Microsoft PhD Fellowship Summary

Applicant: Viska (Sijia) Wei   Sponsor: Prof. Alex Szalay

My proposed research is geared towards designing tools for the next generation AI-telescope powered by deep learning techniques, augmented with insights from astrophysics, computer science and artificial intelligence. The world-wide astronomical community is undertaking a series of comprehensive surveys using this new wide-field telescope, imaging cameras, and software pipelines, with scientific goals that include the structure of our Milky Way Galaxy; the evolution of galaxies; and the distribution and properties of dark matter.

One of the great challenges of designing the survey to address these questions is determining exactly which astronomical objects should be observed in detail (spectroscopically), and how long each should be exposed. As the telescope costs $100k to operate per night, we must carefully select the 500k objects out of billions to target. My research is to develop deep learning algorithms that optimize this target selection with insights from physics, statistics, data science and AI. In particular, I will be developing these tools for the Galactic Archaeology group to understand the makeup of the Milky Way Galaxy and its companions. Mapping the distribution, motions and chemical composition of stars in the major components of our own home galaxy and its neighbor, the Andromeda galaxy, plus smaller satellite systems, will constrain models of galaxy formation, insights that can be directly compared with what we learn from the study of distant galaxies.

My research consists of several interwoven pieces: (1) building generative models for stellar spectra; (2) building physics informed autoencoders; (3) applying these to aid Bayesian Hierarchical Modeling (developed by others); and (4) developing a fast parameter estimation network that can be used as part of the reinforcement learning feedback loop changing target selections.

In order to utilize Deep Learning tools, building sufficient training sets is an essential first step. As there are no comparable observational data in the same spectral range and resolution of PFS, we have to build our networks with simulated magnitudes and spectra. Even the models are not AI-ready, computing a single stellar spectrum with 150,000-pixel resolution takes 15 minutes of CPU time. For the training, several millions of the spectra are needed, which takes 30 years. Training set building can only be feasible if we develop a proxy generative model that create a new spectrum in milliseconds. To accomplish that, we need a highly compressed representation of the available spectral grid, a fast interpolation algorithm, and an autoencoder trained on the interpolated spectra.

Once done, the generated model spectra can be used for multiple purposes. First, they will help in designing a Bayesian hierarchical model that can predict from photometric observations, which of the 500M photometrically observed stars are the best candidates for our project. We can also use the model spectra to train further AI applications, that will measure the physical parameters of a star. These models are expected to perform proper inference and to estimate the uncertainties and covariances of the parameters. This is currently a hot topic.

Based on priors from physics and astrophysics, a hierarchical approach will be developed to improve training accuracy: first partition the data into a moderate number of coarse categories with roughly similar values of the main physical parameters (temperature, surface gravity), then shrink (sparsify) the pixel space to eliminate noise pixels. Starting with statistics and traditional ML based methods, and later moving on to sparse CNNs, we will explore all methods to engineer the best features for deep learning models.

In the second half of my proposed research, as real observational data start to come in, we will modify the autoencoder to learn the differences from the theoretical models and incorporate the incoming new information to improve the predictions of the network. Finally, we can use the neural net to assess the quality of a spectrum after a certain number of exposures and decide whether the accuracy of the prediction is acceptable and we can move to a new target, or we need to observe this object further. We expect that this adoptive strategy with the neural net in a feedback loop for reinforcement learning may yield a significant improvement in the efficiency of collecting adequate high signal-to-noise observations. My research will lead to several refereed publications, already two are in preparation about the data compression and the first experiments with the network design for autoencoders.