Bridging the gap between simulations and survey data - domain adaptation for deep learning in astronomy

Aleksandra Ciprijanovic*[1], Diana Kafkes[1], Kathryn Downey[2], Sydney Jenkins[2], Gabriel Nathan Perdue[1], Sandeep Madireddy[3], Gregory Snyder[4], Brian Nord[1,2,5], and Travis Johnston[6]

[1]Fermi National Accelerator Laboratory – United States
[2]University of Chicago – United States
[3]Argonne national laboratory – United States
[4]Space Telescope Science Institute – United States
[5]Kavli Institute for Cosmological Physics – United States
[6]Oak Ridge National Laboratory – United States

Abstract

Astronomical surveys are already producing very large datasets, and machine learning will play a crucial role in enabling us to fully utilize all of the available data. Machine lerning models are often initially trained on simulated data and then applyed to observations, which can potentially lead to a substantial decrease in model accuracy on the new target dataset. Simulated and telescope data represent different data domains. In order for a machine learning model to work in both domains, domain-invariant learning is necessary. We study the problem of distinguishing between merging and non-merging galaxies in simulated (Illustris-1 cosmological simulation) and observational data (Sloan Digital Sky Survey). Galaxy mergers are very important for our understanding of the evolution of matter in the universe. These are very long processes, so our ability to utilize and combine knowledge from different data domains will be very important for these efforts. In order to unable deep learning algorithms to work in multiple domains we test two domain adaptation techniques: Maximum Mean Discrepancy (MMD) and Domain Adversarial Neural Networks (DANNs). These techniques are particularly important when one of the domains is comprised of new and unlabeled data, which is often the case with new survey data. We show that the addition of domain adaptation improves target domain classification accuracy up to $\sim$20% in the new unlabeled target domain. With further development, these techniques will allow different domain scientists to construct machine learning models that can successfully combine the knowledge from simulated and instrument data or data originating from multiple instruments.

Keywords: machine learning, deep learning, merging galaxies, SDSS, astronomical surveys, domain adaptation, algorithm robustness, Illustris

*Speaker