Self-Supervised ML Adds Depth, Breadth & Speed to Sky Surveys

Scientific Achievement

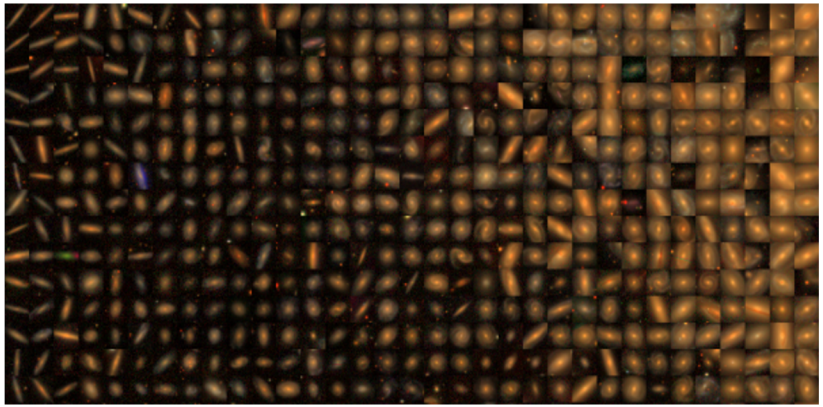
Researchers from Lawrence Berkeley National Laboratory have developed a new self-supervised learning approach that overcomes shortcomings of existing methods for extracting meaningful information from massive sky survey datasets. The method produces representations that can be used to outperform supervised learning methods trained only on labeled data. An early application of this technique was used to discover more than 1,000 previously unknown gravitational lenses.

Significance and Impact

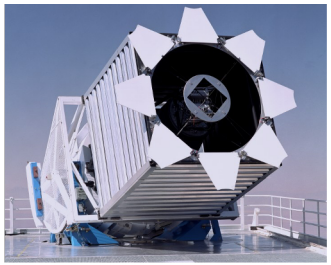
Sky surveys are the largest data generators in astronomy, currently imaging tens of millions to billions of galaxies over the lifetime of a single survey, making automated analysis tools a necessity. The new technique allows for efficient classification of galaxies, distance estimates, similarity searches and outlier detection.

Research Details

The researchers used NERSC's Cori supercomputer to demonstrate their approach. Data preprocessing and preparation was done with Cori's Intel Xeon Phi CPUs, while the bulk of training was conducted using PyTorch on Cori's V100 GPUs. Additional analysis was conducted on Cori's Intel Haswell processors, and the full datasets and source code were made available through NERSC's public data portal.



New techniques are using machine learning to quickly classify and fine tune distances to galaxies. Image: Peter Harrington, Berkeley Lab



The 2.5 meter Sloan Foundation telescope in New Mexico is a major source of survey data.

Hayat, Md Abul; Stein, George; Harrington, Peter; Lukic, Zarija and Mustafa, Mustafa, "Self-supervised Representation Learning for Astronomical Images"; Astrophysical Journal Letters, 911 2021 APR, 10.3847/2041-8213/abf2c7

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC Project PI: Salman Habib, Argonne National Laboratory
DOE Mission Science, Research Funded by Office of High Energy Physics

BERKELEY LAB