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

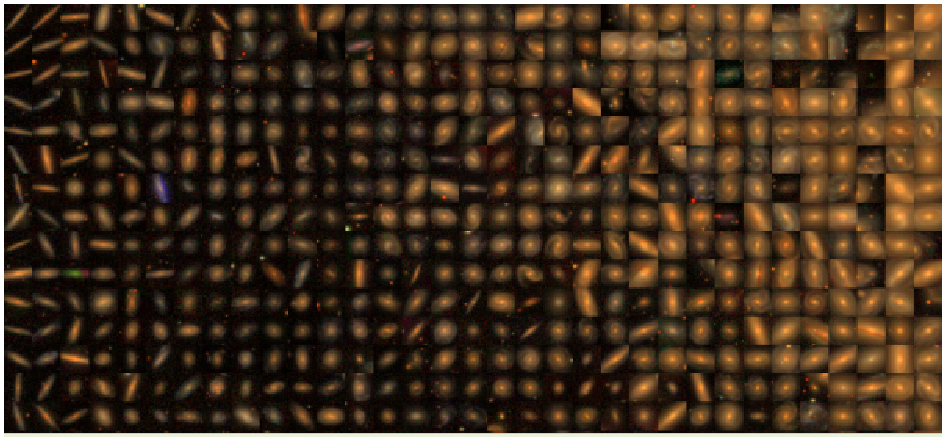
In providing a general map of a particular region, sky surveys are also one of the largest data generators in science

MAY 11, 2021

By Kathy Kincade
Contact: cscomms@lbl.gov

Sky surveys are invaluable for exploring the universe, allowing celestial objects to be catalogued and analyzed without the need for lengthy observations. But in providing a general map or image of a region of the sky, they are also one of the largest data generators in science, currently imaging tens of millions to billions of galaxies over the lifetime of an individual survey. In the near future, for example, the Vera C. Rubin Observatory in Chile will produce 20 TB of data per night, generate about 10 million alerts daily, and end with a final data set of 60 PB in size.

As a result, sky surveys have become increasingly labor-intensive when it comes to sifting through the gathered datasets to find the most relevant information or new discovery. In recent years machine learning has added a welcome twist to the process, primarily in the form of supervised and unsupervised algorithms used to train the computer models that mine the data. But these approaches present their own challenges; for example, supervised learning requires image labels that must be manually assigned, a task that is not only time-consuming but restrictive in scope; at present, only about 1% of all known galaxies have been assigned such labels.



Sky surveys have become increasingly labor-intensive when it comes to sifting through the gathered datasets to find the most relevant information. In recent years machine learning has added a welcome twist to the process. Image: Peter Harrington, Berkeley Lab

To address these limitations, a team of researchers from Lawrence Berkeley National Laboratory (Berkeley Lab) is exploring a new tack: self-supervised representation learning. Like unsupervised learning, self-supervised learning eliminates the need for training labels, instead attempting to learn by comparison. By introducing certain data augmentations, self-supervised algorithms can be used to build "representations" – low-dimensional versions of images that preserve their inherent information – and have recently been demonstrated to outperform supervised learning on industry-standard image datasets.

The Berkeley Lab team presented its research and results in a paper published April 26 in Astrophysical Journal Letters.

"We are quite excited about this work," said George Stein, a post-doctoral researcher at Berkeley Lab and a first author on the new paper. "We believe it is the first to apply state-of-the-art developments in self-supervised learning to large scientific datasets, to great results, and it has already generated a lot of interest from the community."

First author Md Abul Hayat, currently a PhD. student at the University of Arkansas, joined NERSC's summer internship program to collaborate with Mustafa Mustafa when the team began pursuing the idea of applying self-supervised representation learning to sky survey data analysis. Part of their motivation was the growing need to find innovative ways to further automate and speed up the process, given the increasing size of image datasets being produced by the world's ever-more sophisticated telescopes.

"When the Sloan Digital Sky Survey started in the 1990s, it was impossible to do expert labeling on all of their images. Instead, the field moved to crowdsourcing and 'citizen science,' which in turn led to datasets like **Galaxy Zoo**," said Zarija Lukic, a research scientist in Berkeley Lab's Computational Cosmology Center and another co-author on the paper. "But the volume of data that will be coming from the next generation of telescopes is going to be so large that not even crowdsourcing will help you sort out all its images."

"The number of images is increasing day by day, so it has become impossible for a human to go over all of them one by one and provide labels," added Hayat, who continues to work with the Berkeley Lab team on this research. "So eventually the process has to be automated in some way. Our approach is to boil down useful features from these pictures and train the model to come up with a solution from a small part of the data to generalize to an overall representation."

Beyond Sky Surveys

For this proof-of-concept phase of the project, the team applied existing data from ~1.2 million galaxy images generated by the Sloan Digital Sky Survey (SDSS). The goal was to enable the computer model to learn image representations for galaxy morphology classification and photometric redshift estimation, two "downstream" tasks common in sky surveys. In both cases, they found that the self-supervised approach outperformed supervised state-of-the-art results.

"Our approach allows us to learn from the whole sky survey without using any labels, and it can perform a large number of tasks at the same time, each to a higher level of performance than was possible before," Stein noted. "Instead of working to teach a model to do a certain task, you teach it to search all of the data and learn how the images differ from each other, and therefore learn what is in the images themselves.

The idea behind the method is simple to understand, added co-author Peter Harrington, a machine learning engineer at Berkeley Lab. "Given a picture of a galaxy, you can generate different views of it – rotate the galaxy, add a little noise to the image, maybe smear it out with some blurring – and make these little transformations that resemble the noise you have in the telescope itself," he said. "Then you simply teach your model to associate those different views of the same object as similar. That is basically how we build these representations and expose knowledge to the model and make it invariant to the noise."

The research team is now gearing up to apply their approach to a much larger, more complex dataset – the Dark Energy Camera Legacy Survey (DECaLS) – and extend the scope of applications and tasks. Other science areas could benefit from this method as well, Hayat noted, including microscopy, high-energy physics (anomaly detection), medical imaging, and satellite imagery.

"We have demonstrated that self-supervised representation learning on unlabeled data yields notable performance gains over supervised learning for multiple tasks," the research team writes. "The possibility of training a large self-supervised model on massive photometry databases and 'serving' the model for usage by the larger community … is an exciting new direction for machine learning applications in sky surveys."

Just as important, Stein added, "This technique speeds up the science by allowing us to go toward a different way of doing things. It makes it possible for anyone with no machine learning expertise or only small computer power to use it, lowering the barrier to entry to working with these massive datasets."

NERSC is a U.S. Department of Energy Office of Science user facility.

For more information:

NeurIPS Poster: **https://ml4physicalsciences.github.io/2020/files/NeurIPS_ML4PS_2020_17_poster.pdf**

NeurIPS Paper: **https://ml4physicalsciences.github.io/2020/files/NeurIPS_ML4PS_2020_17.pdf**

NERSC video: **https://www.youtube.com/watch?v=LD4Zs8OCrOE**

About Computing Sciences at Berkeley Lab

High performance computing plays a critical role in scientific discovery, and researchers increasingly rely on advances in computer science, mathematics, computational science, data science, and large-scale computing and networking to increase our understanding of ourselves, our planet, and our universe. Berkeley Lab's Computing Sciences Area researches, develops, and deploys new foundations, tools, and technologies to meet these needs and to advance research across a broad range of scientific disciplines.

Founded in 1931 on the belief that the biggest scientific challenges are best addressed by teams, **Lawrence Berkeley National Laboratory** and its scientists have been recognized with 13 Nobel Prizes. Today, Berkeley Lab researchers develop sustainable energy and environmental solutions, create useful new materials, advance the frontiers of computing, and probe the mysteries of life, matter, and the universe. Scientists from around the world rely on the Lab's facilities for their own discovery science. Berkeley Lab is a multiprogram national laboratory, managed by the University of California for the U.S. Department of Energy's Office of Science.

DOE's Office of Science is the single largest supporter of basic research in the physical sciences in the United States, and is working to address some of the most pressing challenges of our time. For more information, please visit **energy.gov/science**.