Stargazing with computers: What machine learning can teach us about the cosmos

February 20 2020, by Shannon Brescher Shea

The Vera Rubin Observatory will house the LSST Camera, which will gather data on 37 billion galaxies and stars over the course of 10 years. Scientists are developing machine learning programs to analyze the flood of data. Credit: M. Park/Inigo Films/LSST/AURA/NSF

Gazing up at the night sky in a rural area, you'll probably see the shining

moon surrounded by stars. If you're lucky, you might spot the furthest thing visible with the naked eye—the Andromeda galaxy. It's the nearest neighbor to our galaxy, the Milky Way. But that's just the tiniest fraction of what's out there. When the Department of Energy's (DOE) Legacy Survey of Space and Time (LSST) Camera at the National Science Foundation's Vera Rubin Observatory turns on in 2022, it will take photos of 37 billion galaxies and stars over the course of a decade.

The output from this huge telescope will swamp researchers with data. In those 10 years, the LSST Camera will take 2,000 photos for each patch of the Southern Sky it covers. Each image can have up to a million objects in it.

"In terms of the scale of the data, the amount of the data, the complexity of the data, they're well beyond any of the current data sets we have," said Rachel Mandelbaum, a professor at Carnegie Mellon University and LSST Dark Energy Science Collaboration spokesperson. "This opens up a huge amount of discovery space."

Scientists aren't building the LSST Camera to just take pretty pictures. They want to identify, categorize, and measure celestial objects that can reveal information about the very structure of the universe. Understanding dark energy and other cosmological mysteries requires data on supernovae and galaxies. Researchers may even find entirely new classes of objects.

"There are going to be some objects that we have never seen before because that is the point of new discovery," said Renée Hložek, an assistant professor of astrophysics at the University of Toronto, who works with the LSST Dark Energy Science Collaboration. "We will find a bunch of what we call weirdos, or anomalies."

The sheer volume and strangeness of the data will make it difficult to

analyze. While a stargazer new to an area might go out in the field with a local expert, scientists don't have such a guide to new pieces of the universe. So they're making their own. More accurately, they're making many different guides that can help them identify and categorize these objects. Astrophysicists supported by the DOE Office of Science are developing these guides in the form of computer models that rely on machine learning to examine the LSST data. Machine learning is a process where a computer program learns over time about the relationships in a set of data.

Computer Programs that Learn

Processing data quickly is a must for scientists in the Dark Energy Science Collaboration. Scientists need to know that the camera is pointing at exactly the right place and taking data correctly each time. This quick processing also helps them know if anything has changed in that part of the sky since the last time they took photos of it. Subtracting the current photo from previous ones shows them if there's a sign of an interesting celestial object or phenomenon.

They also need to combine a lot of photos together in a way that's accurate and usable. This project is looking into the depths of the universe to capture images of some of the faintest stars and galaxies. It will also be taking photos in less-than-ideal atmospheric conditions. To compensate, scientists need programs that can combine images together to improve clarity.

Machine learning can tackle these challenges in addition to handling the sheer amount of data. As these programs analyze more data, the more accurate they become. Just like a person learning to identify a constellation, they gain better judgment over time.

"Many scientists regard machine learning as the most promising option

for classifying sources based on photometric measurements (measurements of light intensity)," said Eve Kovacs, a physicist at DOE's Argonne National Laboratory.

But machine learning programs need to teach themselves before they can tackle a pile of new data. There are two main ways to "train" a machine learning program: unsupervised and supervised.

Unsupervised machine learning is like someone teaching themselves about stars from just their nightly observations. The program trains itself on unlabeled data. While unsupervised machine learning can group images and identify outliers, it can't categorize them without a guidebook of some sort.

Supervised machine learning is like a newbie relying on a guidebook. The researchers feed it a massive set of data that is labeled with the classes of each object. By examining the data over and over, the program learns the relationship between the observation and the labels. This technique is especially useful for classifying objects into known groups.

In some cases, the researchers also feed the program a specific set of features to look for, like brightness, shape, or color. They provide guidance on how important each feature is compared to the others. In other programs, the machine learning program figures out the relevant features itself.

However, the accuracy of supervised machine learning depends on having a good training set, with all of the diversity and variability of a real one. For photos from the LSST Camera, that variability could include streaks from satellites moving across the sky. The labeling also has to be extremely accurate.

"We have to put as much physics as we can into the training sets," said

Mandelbaum. "It doesn't remove from us the burden to understand the physics. It just moves it into a different part of the problem."

Mile Markers on the Space Highway

Some of the universe's most interesting objects don't stick around for long. Transient objects appear very bright, fade over a specific period of time, and then go dark. Supernovae—massively exploding stars—are one kind of transient object. Variable objects change in brightness over time in a consistent way. Certain types of both can be "standard candles," items scientists can use to measure distance from Earth, like mile markers on an interstate. These standard candles provide information about the universe's size and history.

"If you look at enough galaxies on a given night, you're almost guaranteed to discover a supernova," said Kovacs.

To know if a supernova is going to be useful as a standard candle or not, scientists need to know what type it is. Type Ia supernovae can be standard candles. Just like drawing on experience can tell stargazers if they're looking at Mars or Venus, a computer program can use its training to classify a supernova from an image.

"The little fly in the ointment in all of this is that the Type Ia supernovae aren't exactly standard candles. They have a certain amount of variation," said Kovacs. "Understanding that variation … actually lies at the heart of making all of this work."

Kovacs and her collaborators created a program that uses supernovae's colors to sort them into categories. Previously, scientists trained machine learning algorithms by having them compare a specific supernova's brightness over time to a model based on Type Ia supernova. But the programs were likely to misclassify too many supernovae as Type Ia.

Her team took a different approach. They identified a set of 17 features characterizing the light curves (time variation of light intensity) of supernovae. Using a training set of several thousand simulated supernovae, they were able to achieve classifications that had extremely high levels of accuracy.

Figuring out how far cosmic objects are from Earth is another promising area for machine learning. Previously, scientists relied on spectroscopic telescopes that use fiber optics to precisely measure these objects' distances. But the LSST Camera is going to find more than 1,000 transient objects a night. That's too many to follow up on using this technique. Mandelbaum and her team developed a machine learning program that can estimate this distance accurately from photos alone. It can also adapt and incorporate spectroscopic data if it's available.

But supernovae aren't the only objects that can be used as standard candles. In fact, astrophysicists often use other objects to calibrate their distance. Mandelbaum and her team used machine learning to find other potential standard candles. By feeding the program data about lots of variable stars, they found that it could come up with and apply features that identify a good standard candle without needing to classify the star first. Skipping that step—which requires a lot of labeled, categorized data—simplified the process. It also helped avoid biases or errors from classification. The program produced a sample with stars that were just as good standard candles as Cepheids, a useful but rare variable star. There was another bonus—the stars in their sample were generally brighter and easier to measure than Cepheids.

"The machine learning helps you ferret out these complicated spaces because humans have difficulty thinking in more than three dimensions," Kovacs said.

Picking and Choosing at a Galactic Level

While individual stars can reveal a great deal of information, sometimes you need a whole galaxy. Using a photo alone, it's easier to figure out the distance of the host galaxy of a supernova rather than the supernova itself. But scientists must pick the right host galaxy. In the past, they've done this matching by hand. But the LSST Camera is going to create way too much data for humans to handle.

In one of Kovac's projects, the scientific team developed an algorithm that matched the host galaxy to the supernova correctly 90 to 92 percent of the time. Not accurate enough. But machine learning came to the rescue. The team developed a machine learning program to tell them how likely any classification was to be right or wrong. It identified seven to eight percent of the original output as most likely wrong. Removing those items from the data increased the accuracy and made it easier to follow up on the tricky photos by hand.

Tapping the Collective Mind

To further explore the power of machine learning, two of the LSST Camera's science groups found a unique way to draw on scientists' brainpower—they ran a contest. Partnering with Kaggle, a website for data scientists, they targeted non-astronomers specializing in machine learning to develop programs to sort through future data from the LSST Camera.

"If you only speak to the people you know, you lose that diversity of thought of the larger community," said Hložek, who ran the competition. "We wanted folks to actually work together to pool their models and pool their data."

They particularly wanted the programs to pick out object types that astrophysicists may not have seen before. They gave the group three

million objects to sort into 15 categories, with the 15th being 'I haven't seen it before.'

"We want to prime ourselves to be open to that kind of work," Hložek said. "What are the ways that weirdness can manifest?"

More than 1,300 competitors in 1,000 teams participated in the challenge, which ended in December 2018. Now, researchers on the LSST Camera are sorting through the codes to combine them into the best possible set of programs.

All of this activity is happening years before the LSST Camera even gets turned on. Machine learning programs are sure to reveal even more once the data starts flowing in. While computers can't gaze at the stars in wonder, they'll provide ever more insight into the celestial objects that inspire such awe in us.

More information: For more information, please visit www.energy.gov/science.

Provided by Argonne National Laboratory

Citation: Stargazing with computers: What machine learning can teach us about the cosmos (2020, February 20) retrieved 17 June 2024 from https://phys.org/news/2020-02-stargazing-machine-cosmos.html