Using generative AI to improve software testing

March 5 2024, by Zach Winn

DataCebo offers a generative software system called the Synthetic Data Vault to help organizations create synthetic data to do things like test software applications and train machine learning models. Credit: DataCebo, edited by MIT News

Generative AI is getting plenty of attention for its ability to create text

and images. But those media represent only a fraction of the data that proliferate in our society today. Data are generated every time a patient goes through a medical system, a storm impacts a flight, or a person interacts with a software application.

Using generative AI to create realistic synthetic data around those scenarios can help organizations more effectively treat patients, reroute planes, or improve software platforms—especially in scenarios where real-world data are limited or sensitive.

For the last three years, the MIT spinout DataCebo has offered a generative software system called the Synthetic Data Vault to help organizations create synthetic data to do things like test software applications and train machine learning models.

The Synthetic Data Vault, or SDV, has been downloaded more than 1 million times, with more than 10,000 data scientists using the open-source library for generating synthetic tabular data. The founders—Principal Research Scientist Kalyan Veeramachaneni and alumna Neha Patki '15, SM '16—believe the company's success is due to SDV's ability to revolutionize software testing.

SDV goes viral

In 2016, Veeramachaneni's group in the Data to AI Lab unveiled a suite of open-source generative AI tools to help organizations create synthetic data that matched the statistical properties of real data.

Companies can use synthetic data instead of sensitive information in programs while still preserving the statistical relationships between datapoints. Companies can also use synthetic data to run new software through simulations to see how it performs before releasing it to the public.

Veeramachaneni's group came across the problem because it was working with companies that wanted to share their data for research.

"MIT helps you see all these different use cases," Patki explains. "You work with finance companies and health care companies, and all those projects are useful to formulate solutions across industries."

In 2020, the researchers founded DataCebo to build more SDV features for larger organizations. Since then, the use cases have been as impressive as they've been varied.

With DataCebo's new flight simulator, for instance, airlines can plan for rare weather events in a way that would be impossible using only historic data. In another application, SDV users synthesized medical records to predict health outcomes for patients with cystic fibrosis. A team from Norway recently used SDV to create synthetic student data to evaluate whether various admissions policies were meritocratic and free from bias.

In 2021, the data science platform Kaggle hosted a competition for data scientists that used SDV to create synthetic data sets to avoid using proprietary data. Roughly 30,000 data scientists participated, building solutions and predicting outcomes based on the company's realistic data.

And as DataCebo has grown, it's stayed true to its MIT roots: All of the company's current employees are MIT alumni.

Supercharging software testing

Although their open-source tools are being used for a variety of use cases, the company is focused on growing its traction in software testing.

"You need data to test these software applications," Veeramachaneni

says. "Traditionally, developers manually write scripts to create synthetic data. With generative models, created using SDV, you can learn from a sample of data collected and then sample a large volume of synthetic data (which has the same properties as real data), or create specific scenarios and edge cases, and use the data to test your application."

For example, if a bank wanted to test a program designed to reject transfers from accounts with no money in them, it would have to simulate many accounts simultaneously transacting. Doing that with data created manually would take a lot of time. With DataCebo's generative models, customers can create any edge case they want to test.

"It's common for industries to have data that is sensitive in some capacity," Patki says. "Often when you're in a domain with sensitive data you're dealing with regulations, and even if there aren't legal regulations, it's in companies' best interest to be diligent about who gets access to what at which time. So, synthetic data is always better from a privacy perspective."

Scaling synthetic data

Veeramachaneni believes DataCebo is advancing the field of what it calls synthetic enterprise data, or data generated from user behavior on large companies' software applications.

"Enterprise data of this kind is complex, and there is no universal availability of it, unlike language data," Veeramachaneni says. "When folks use our publicly available software and report back if works on a certain pattern, we learn a lot of these unique patterns, and it allows us to improve our algorithms. From one perspective, we are building a corpus of these complex patterns, which for language and images is readily available. "

DataCebo also recently released features to improve SDV's usefulness, including tools to assess the "realism" of the generated data, called the SDMetrics library as well as a way to compare models' performances called SDGym.

"It's about ensuring organizations trust this new data," Veeramachaneni says. "[Our tools offer] programmable synthetic data, which means we allow enterprises to insert their specific insight and intuition to build more transparent models."

As companies in every industry rush to adopt AI and other data science tools, DataCebo is ultimately helping them do so in a way that is more transparent and responsible.

"In the next few years, synthetic data from generative models will transform all data work," Veeramachaneni says. "We believe 90% of enterprise operations can be done with synthetic data."

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Using generative AI to improve software testing (2024, March 5) retrieved 28 April 2024 from https://techxplore.com/news/2024-03-generative-ai-software.html