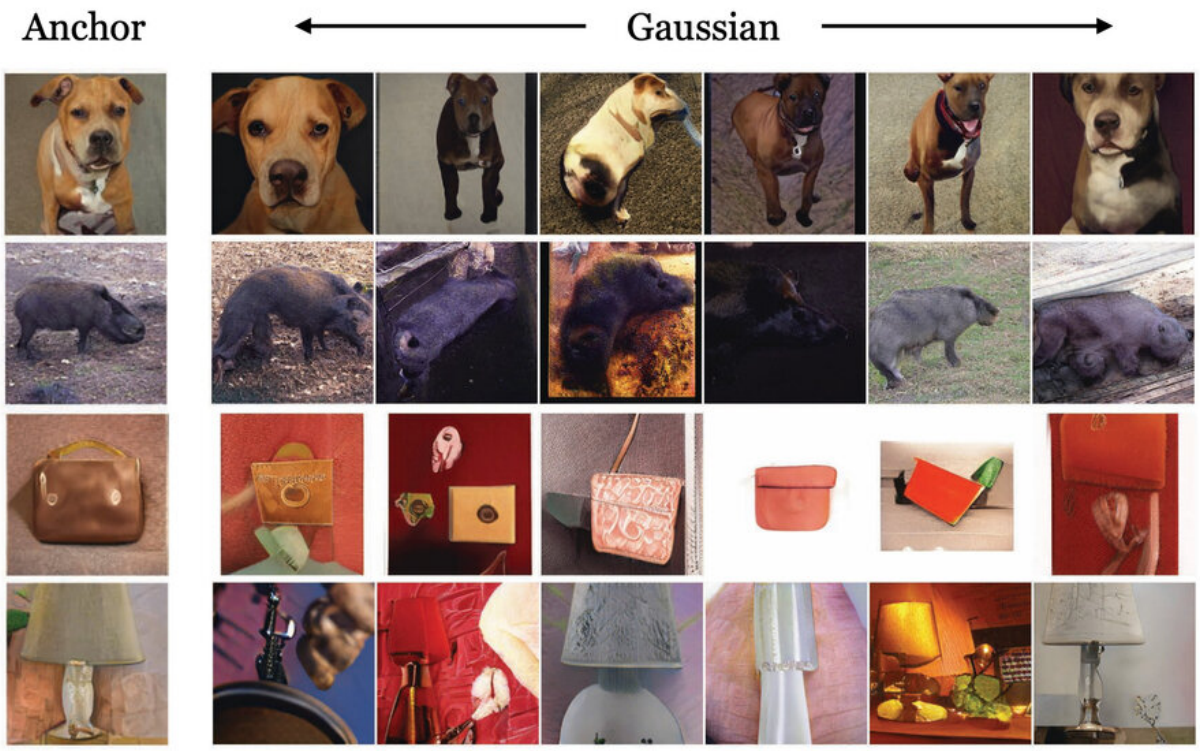
When it comes to AI, can we ditch the datasets?

March 15 2022, by Adam Zewe



MIT researchers have demonstrated the use of a generative machine-learning model to create synthetic data, based on real data, that can be used to train another model for image classification. This image shows examples of the generative model's transformation methods. Credit: Massachusetts Institute of Technology

Huge amounts of data are needed to train machine-learning models to perform image classification tasks, such as identifying damage in satellite photos following a natural disaster. However, these data are not always easy to come by. Datasets may cost millions of dollars to generate, if usable data exist in the first place, and even the best datasets often contain biases that negatively impact a model's performance.

To circumvent some of the problems presented by datasets, MIT researchers developed a method for training a machine learning model that, rather than using a dataset, uses a special type of machine-learning model to generate extremely realistic synthetic data that can train another model for downstream vision tasks.

Their results show that a contrastive representation learning model trained using only these synthetic data is able to learn visual representations that rival or even outperform those learned from real data.

This special machine-learning model, known as a generative model, requires far less memory to store or share than a dataset. Using synthetic data also has the potential to sidestep some concerns around privacy and usage rights that limit how some real data can be distributed. A generative model could also be edited to remove certain attributes, like race or gender, which could address some biases that exist in traditional datasets.

"We knew that this method should eventually work; we just needed to wait for these generative models to get better and better. But we were especially pleased when we showed that this method sometimes does even better than the real thing," says Ali Jahanian, a research scientist in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and lead author of the paper.

Jahanian wrote the paper with CSAIL grad students Xavier Puig and Yonglong Tian, and senior author Phillip Isola, an assistant professor in the Department of Electrical Engineering and Computer Science. The research will be presented at the International Conference on Learning Representations.

Generating synthetic data

Once a generative model has been trained on real data, it can generate synthetic data that are so realistic they are nearly indistinguishable from the real thing. The training process involves showing the generative model millions of images that contain objects in a particular class (like cars or cats), and then it learns what a car or cat looks like so it can generate similar objects.

Essentially by flipping a switch, researchers can use a pretrained generative model to output a steady stream of unique, realistic images that are based on those in the model's training dataset, Jahanian says.

But generative models are even more useful because they learn how to transform the underlying data on which they are trained, he says. If the model is trained on images of cars, it can "imagine" how a car would look in different situations—situations it did not see during training—and then output images that show the car in unique poses, colors, or sizes.

Having multiple views of the same image is important for a technique called contrastive learning, where a machine-learning model is shown many unlabeled images to learn which pairs are similar or different.

The researchers connected a pretrained generative model to a contrastive learning model in a way that allowed the two models to work together automatically. The contrastive learner could tell the generative model to

produce different views of an object, and then learn to identify that object from multiple angles, Jahanian explains.

"This was like connecting two building blocks. Because the generative model can give us different views of the same thing, it can help the contrastive method to learn better representations," he says.

Even better than the real thing

The researchers compared their method to several other image classification models that were trained using real data and found that their method performed as well, and sometimes better, than the other models.

One advantage of using a generative model is that it can, in theory, create an infinite number of samples. So, the researchers also studied how the number of samples influenced the model's performance. They found that, in some instances, generating larger numbers of unique samples led to additional improvements.

"The cool thing about these generative models is that someone else trained them for you. You can find them in online repositories, so everyone can use them. And you don't need to intervene in the model to get good representations," Jahanian says.

But he cautions that there are some limitations to using generative models. In some cases, these models can reveal source data, which can pose privacy risks, and they could amplify biases in the datasets they are trained on if they aren't properly audited.

He and his collaborators plan to address those limitations in future work. Another area they want to explore is using this technique to generate corner cases that could improve machine learning models. Corner cases

often can't be learned from real data. For instance, if researchers are training a computer vision model for a self-driving car, real data wouldn't contain examples of a dog and his owner running down a highway, so the model would never learn what to do in this situation. Generating that corner case data synthetically could improve the performance of machine learning models in some high-stakes situations.

The researchers also want to continue improving generative models so they can compose images that are even more sophisticated, he says.

More information: Paper: Generative models as a data source for multiview representation learning.
openreview.net/pdf?id=qhAeZjs7dCL

Provided by Massachusetts Institute of Technology