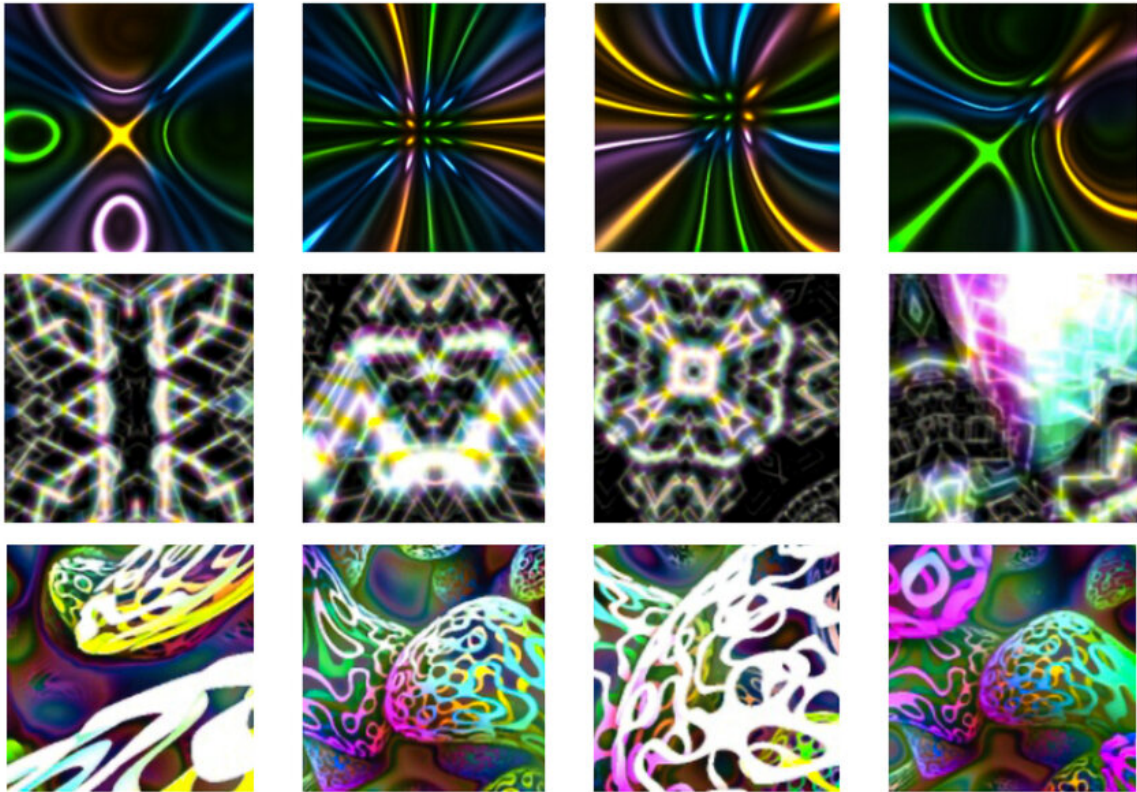# A simpler path to better computer vision

November 23 2022, by Adam Zewe

Researchers used a large collection of simple, un-curated synthetic image generation programs to pretrain a computer vision model for image classification. The researchers didn't curate or alter the programs, which each comprised just a few lines of code. In this image, the image sets in each row were produced using 3 different image generation programs. Credit: Massachusetts Institute of Technology

Before a machine-learning model can complete a task, such as identifying cancer in medical images, the model must be trained. Training image classification models typically involves showing the model millions of example images gathered into a massive dataset.

However, using real image data can raise practical and [ethical concerns](): The images could run afoul of copyright laws, violate people's privacy, or be biased against a certain racial or ethnic group. To avoid these pitfalls, researchers can use image generation programs to create [synthetic data]() for model training. But these techniques are limited because expert knowledge is often needed to hand-design an image generation program that can create effective training data.

Researchers from MIT, the MIT-IBM Watson AI Lab, and elsewhere took a different approach. Instead of designing customized image generation programs for a particular training task, they gathered a dataset of 21,000 publicly available programs from the internet. Then they used this large collection of basic image generation programs to train a computer vision model.

These programs produce diverse images that display simple colors and textures. The researchers didn't curate or alter the programs, which each comprised just a few lines of code.

The models they trained with this large dataset of programs classified images more accurately than other synthetically trained models. And, while their models underperformed those trained with real data, the researchers showed that increasing the number of image programs in the dataset also increased model performance, revealing a path to attaining higher accuracy.

"It turns out that using lots of programs that are uncurated is actually better than using a small set of programs that people need to manipulate.

Data are important, but we have shown that you can go pretty far without real data," says Manel Baradad, an [electrical engineering](link) and computer science (EECS) graduate student working in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and lead author of the paper describing this technique.

Co-authors include Tongzhou Wang, an EECS grad student in CSAIL; Rogerio Feris, principal scientist and manager at the MIT-IBM Watson AI Lab; Antonio Torralba, the Delta Electronics Professor of Electrical Engineering and Computer Science and a member of CSAIL; and senior author Phillip Isola, an associate professor in EECS and CSAIL; along with others at JPMorgan Chase Bank and Xyla, Inc. The research will be presented at the Conference on Neural Information Processing Systems.

Rethinking pretraining

Machine-learning models are typically pretrained, which means they are trained on one dataset first to help them build parameters that can be used to tackle a different task. A model for classifying X-rays might be pretrained using a huge dataset of synthetically generated images before it is trained for its actual task using a much smaller dataset of real X-rays.

These researchers previously showed that they could use a handful of image generation programs to create synthetic data for model pretraining, but the programs needed to be carefully designed so the synthetic images matched up with certain properties of real images. This made the technique difficult to scale up.

In the new work, they used an enormous dataset of uncurated image generation programs instead.

They began by gathering a collection of 21,000 images generation

programs from the internet. All the programs are written in a simple programming language and comprise just a few snippets of code, so they generate images rapidly.

"These programs have been designed by developers all over the world to produce images that have some of the properties we are interested in. They produce images that look kind of like abstract art," Baradad explains.

These simple programs can run so quickly that the researchers didn't need to produce images in advance to train the model. The researchers found they could generate images and train the model simultaneously, which streamlines the process.

They used their massive dataset of image generation programs to pretrain computer vision models for both supervised and unsupervised image classification tasks. In supervised learning, the image data are labeled, while in unsupervised learning the model learns to categorize images without labels.

Improving accuracy

When they compared their pretrained models to state-of-the-art computer vision models that had been pretrained using synthetic data, their models were more accurate, meaning they put images into the correct categories more often. While the accuracy levels were still less than models trained on real data, their technique narrowed the performance gap between models trained on real data and those trained on synthetic data by 38 percent.

"Importantly, we show that for the number of programs you collect, performance scales logarithmically. We do not saturate performance, so if we collect more programs, the model would perform even better. So,

there is a way to extend our approach," Manel says.

The researchers also used each individual image generation program for pretraining, in an effort to uncover factors that contribute to model accuracy. They found that when a program generates a more diverse set of images, the model performs better. They also found that colorful images with scenes that fill the entire canvas tend to improve model performance the most.

Now that they have demonstrated the success of this pretraining approach, the researchers want to extend their technique to other types of data, such as multimodal data that include text and images. They also want to continue exploring ways to improve image classification performance.

"There is still a gap to close with models trained on real data. This gives our research a direction that we hope others will follow," he says.

More information: Conference: nips.cc/

Paper: openreview.net/pdf?id=wJwHTgIoE0P

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