In machine learning, synthetic data can offer real performance improvements

BY ADAM ZEWE

Teaching a machine to recognize human actions has many potential applications, such as automatically detecting workers who fall at a construction site or enabling a smart home robot to interpret a user's gestures.

Teaching a machine to recognize human actions has many potential applications, such as automatically detecting workers who fall at a construction site or enabling a smart home robot to interpret a user's gestures.

To do this, researchers train machine-learning models using vast datasets of video clips that show humans performing actions. However, not only is it expensive and laborious to gather and label millions or billions of videos, but the clips often contain sensitive information, like people's faces or license plate numbers. Using these videos might also violate copyright or data protection laws. And this assumes the video data are publicly available in the first place — many datasets are owned by companies and aren't free to use.

So, researchers are turning to synthetic datasets. These are made by a computer that uses 3D models of scenes, objects, and humans to quickly produce many varying clips of specific actions — without the potential copyright issues or ethical concerns that come with real data.

But are synthetic data as "good" as real data? How well does a model trained with these data perform when it's asked to classify real human actions? A team of researchers at MIT, the MIT-IBM Watson AI Lab, and Boston University sought to answer this question. They built a synthetic dataset of 150,000 video clips that captured a wide range of human actions, which they used to train machine-learning models. Then they showed these models six datasets of real-world videos to see how well they could learn to recognize actions in those clips.

The researchers found that the synthetically trained models performed even better than models trained on real data for videos that have fewer background objects.

This work could help researchers use synthetic datasets in such a way that models achieve higher accuracy on real-world tasks. It could also help scientists identify which machine-learning applications could be best suited for training with synthetic data, to mitigate some of the ethical, privacy, and copyright concerns of using real datasets.

"The ultimate goal of our research is to replace real data pretraining with synthetic data pretraining. There is a cost in creating an action in synthetic data, but once that is done, then you can generate an unlimited number of images or videos by changing the pose, the lighting, etc. That is the beauty of synthetic data," says Rogerio Feris, principal scientist and manager at the MIT-IBM Watson AI Lab, and co-author of a paper detailing this research.

The paper is authored by lead author Yo-whan "John" Kim '22; Aude Oliva, director of strategic industry engagement at the MIT Schwarzman College of Computing, MIT director of the MIT-

IBM Watson AI Lab, and a senior research scientist in the Computer Science and Artificial Intelligence Laboratory (CSAIL); and seven others. The research will be presented at the Conference on Neural Information Processing Systems.

Building a synthetic dataset

The researchers began by compiling a new dataset using three publicly available datasets of synthetic video clips that captured human actions. Their dataset, called Synthetic Action Pre-training and Transfer (SynAPT), contained 150 action categories, with 1,000 video clips per category.

They selected as many action categories as possible, such as people waving or falling on the floor, depending on the availability of clips that contained clean video data.

Once the dataset was prepared, they used it to pretrain three machine-learning models to recognize the actions. Pretraining involves training a model for one task to give it a head-start for learning other tasks. Inspired by the way people learn — we reuse old knowledge when we learn something new — the pretrained model can use the parameters it has already learned to help it learn a new task with a new dataset faster and more effectively.

They tested the pretrained models using six datasets of real video clips, each capturing classes of actions that were different from those in the training data.

It surprised the researchers to see that all three synthetic models outperformed models trained with real video clips on four of the six datasets. Their accuracy was highest for datasets that contained video clips with “low scene-object bias.”

Low scene-object bias means that the model cannot recognize the action by looking at the background or other objects in the scene — it must focus on the action itself. For example, if the model is tasked with classifying diving poses in video clips of people diving into a swimming pool, it cannot identify a pose by looking at the water or the tiles on the wall. It must focus on the person’s motion and position to classify the action.

“In videos with low scene-object bias, the temporal dynamics of the actions is more important than the appearance of the objects or the background, and that seems to be well-captured with synthetic data,” Feris says.

“High scene-object bias can actually act as an obstacle. The model might misclassify an action by looking at an object, not the action itself. It can confuse the model,” Kim explains.

Boosting performance

Building off these results, the researchers want to include more action classes and additional synthetic video platforms in future work, eventually creating a catalog of models that have been pre-trained using synthetic data, says co-author Rameswar Panda, a research staff member at the MIT-IBM Watson AI Lab.

“We want to build models which have very similar performance or even better performance than the existing models in the literature, but without being bound by any of those biases or security concerns,” he adds.

They also want to combine their work with research that seeks to generate more accurate and realistic synthetic videos, which could boost the performance of the models, says SouYoung Jin, a co-author and CSAIL postdoc. She is also interested in exploring how models might learn differently when they are trained with synthetic data.

“We use synthetic datasets to prevent privacy issues or contextual or social bias, but what does the model actually learn? Does it learn something that is unbiased?” she says.

Now that they have demonstrated this use potential for synthetic videos, they hope other researchers will build upon their work.

“Despite there being a lower cost to obtaining well-annotated synthetic data, currently we do not have a dataset with the scale to rival the biggest annotated datasets with real videos. By discussing the different costs and concerns with real videos, and showing the efficacy of synthetic data, we hope to motivate efforts in this direction,” adds co-author Samarth Mishra, a graduate student at Boston University (BU).

Additional co-authors include Hilde Kuehne, professor of computer science at Goethe University in Germany and an affiliated professor at the MIT-IBM Watson AI Lab; Leonid Karlinsky, research staff member at the MIT-IBM Watson AI Lab; Venkatesh Saligrama, professor in the Department of Electrical and Computer Engineering at BU; and Kate Saenko, associate professor in the Department of Computer Science at BU and a consulting professor at the MIT-IBM Watson AI Lab.

This research was supported by the Defense Advanced Research Projects Agency LwLL, as well as the MIT-IBM Watson AI Lab and its member companies, Nexplore and Woodside.