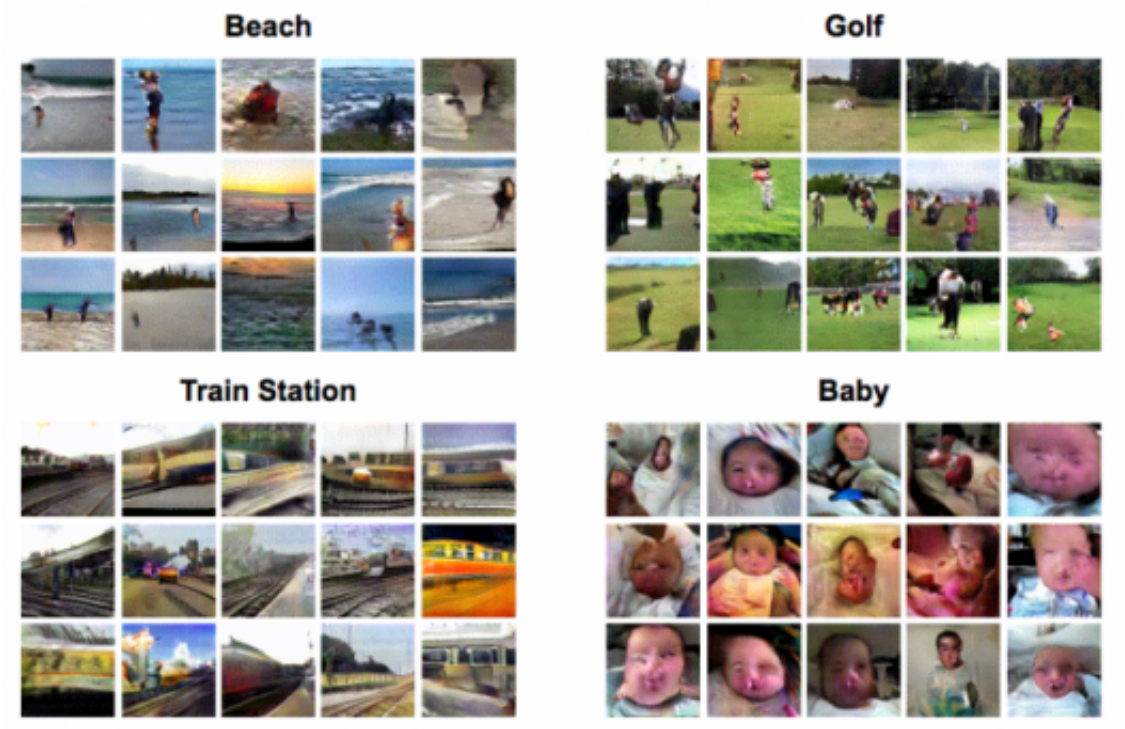
Deep-learning algorithm creates videos of the future

November 29 2016, by Adam Conner-Simons



Given a still image from a scene, the CSAIL team's deep-learning algorithm can create a brief video that simulates the future of that scene. Credit: Massachusetts Institute of Technology

Living in a dynamic physical world, it's easy to forget how effortlessly we understand our surroundings. With minimal thought, we can figure out how scenes change and objects interact.

But what's second nature for us is still a huge problem for machines. With the limitless number of ways that objects can move, teaching computers to predict future actions can be difficult.

Recently, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have moved a step closer, developing a deep-learning algorithm that, given a still image from a scene, can create a brief video that simulates the future of that scene.

Trained on 2 million unlabeled videos that include a year's worth of footage, the algorithm generated videos that human subjects deemed to be realistic 20 percent more often than a baseline model.

The team says that future versions could be used for everything from improved security tactics and safer self-driving cars. According to CSAIL PhD student and first author Carl Vondrick, the algorithm can also help machines recognize people's activities without expensive human annotations.

"These videos show us what computers think can happen in a scene," says Vondrick. "If you can predict the future, you must have understood something about the present."

Vondrick wrote the paper with MIT professor Antonio Torralba and Hamed Pirsiavash, a former CSAIL postdoc who is now a professor at the University of Maryland Baltimore County (UMBC). The work will be presented at next week's Neural Information Processing Systems (NIPS) conference in Barcelona.

How it works

Multiple researchers have tackled similar topics in computer vision, including MIT Professor Bill Freeman, whose new work on "visual

dynamics" also creates future frames in a scene. But where his model focuses on extrapolating videos into the future, Torralba's model can also generate completely new videos that haven't been seen before.

Previous systems build up scenes frame by frame, which creates a large margin for error. In contrast, this work focuses on processing the entire scene at once, with the algorithm generating as many as 32 frames from scratch per second.

"Building up a scene frame-by-frame is like a big game of 'Telephone,' which means that the message falls apart by the time you go around the whole room," says Vondrick. "By instead trying to predict all frames simultaneously, it's as if you're talking to everyone in the room at once."

Of course, there's a trade-off to generating all frames simultaneously: While it becomes more accurate, the computer model also becomes more complex for longer videos. Nevertheless, this complexity may be worth it for sharper predictions.

To create multiple frames, researchers taught the model to generate the foreground separate from the background, and to then place the objects in the scene to let the model learn which objects move and which objects don't.

The team used a deep-learning method called "adversarial learning" that involves training two competing neural networks. One network generates video, and the other discriminates between the real and generated videos. Over time, the generator learns to fool the discriminator.

From that, the model can create videos resembling scenes from beaches, train stations, hospitals, and golf courses. For example, the beach model produces beaches with crashing waves, and the golf model has people walking on grass.

Testing the scene

The team compared the videos against a baseline of generated videos and asked subjects which they thought were more realistic. From over 13,000 opinions of 150 users, subjects chose the generative model videos 20 percent more often than the baseline.

Vondrick stresses that the model still lacks some fairly simple common-sense principles. For example, it often doesn't understand that objects are still there when they move, like when a train passes through a scene. The model also tends to make humans and objects look much larger in size than reality.

Another limitation is that the generated videos are just one and a half seconds long, which the team hopes to be able to increase in future work. The challenge is that this requires tracking longer dependencies to ensure that the scene still makes sense over longer time periods. One way to do this would be to add human supervision.

"It's difficult to aggregate accurate information across long time periods in videos," says Vondrick. "If the video has both cooking and eating activities, you have to be able to link those two together to make sense of the scene."

These types of models aren't limited to predicting the future. Generative videos can be used for adding animation to still images, like the animated newspaper from the Harry Potter books. They could also help detect anomalies in security footage and compress data for storing and sending longer videos.

"In the future, this will let us scale up vision systems to recognize objects and scenes without any supervision, simply by training them on video," says Vondrick.

More information: Paper: "Generating Videos with Scene Dynamics" web.mit.edu/vondrick/tinyvideo/paper.pdf

Provided by Massachusetts Institute of Technology