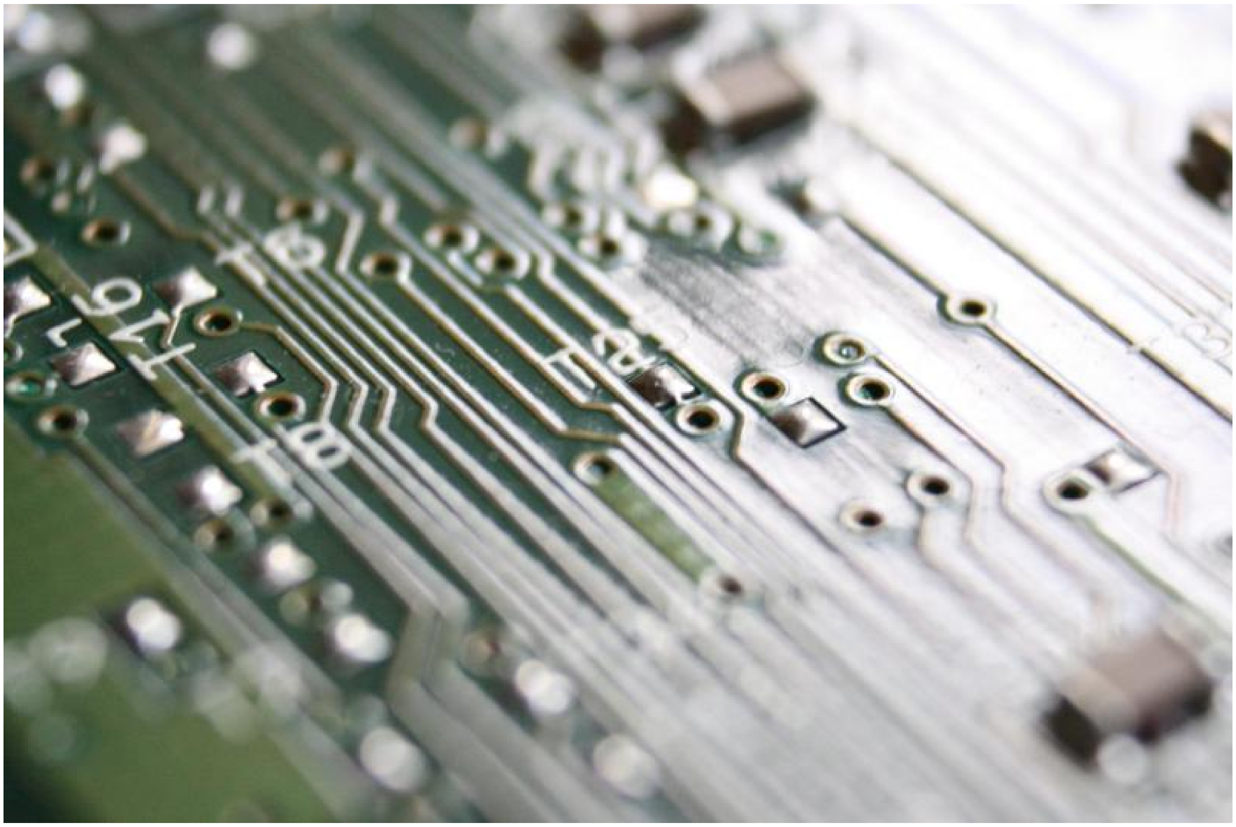
Computer learns to recognize sounds by watching video

December 1 2016



Credit: Public Domain

In recent years, computers have gotten remarkably good at recognizing speech and images: Think of the dictation software on most cellphones, or the algorithms that automatically identify people in photos posted to

Facebook.

But recognition of natural sounds—such as crowds cheering or waves crashing—has lagged behind. That's because most automated recognition systems, whether they process audio or visual information, are the result of machine learning, in which computers search for patterns in huge compendia of training data. Usually, the training data has to be first annotated by hand, which is prohibitively expensive for all but the highest-demand applications.

Sound recognition may be catching up, however, thanks to researchers at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL). At the Neural Information Processing Systems conference next week, they will present a sound-recognition system that outperforms its predecessors but didn't require hand-annotated data during training.

Instead, the researchers trained the system on video. First, existing computer vision systems that recognize scenes and objects categorized the images in the video. The new system then found correlations between those visual categories and natural sounds.

"Computer vision has gotten so good that we can transfer it to other domains," says Carl Vondrick, an MIT graduate student in electrical engineering and computer science and one of the paper's two first authors. "We're capitalizing on the natural synchronization between vision and sound. We scale up with tons of unlabeled video to learn to understand sound."

The researchers tested their system on two standard databases of annotated sound recordings, and it was between 13 and 15 percent more accurate than the best-performing previous system. On a data set with 10 different sound categories, it could categorize sounds with 92 percent accuracy, and on a data set with 50 categories it performed with 74 percent accuracy. On those same data sets, humans are 96 percent and 81 percent accurate, respectively.

"Even humans are ambiguous," says Yusuf Aytar, the paper's other first author and a postdoc in the lab of MIT professor of electrical engineering and computer science Antonio Torralba. Torralba is the final co-author on the paper.

"We did an experiment with Carl," Aytar says. "Carl was looking at the computer monitor, and I couldn't see it. He would play a recording and I would try to guess what it was. It turns out this is really, really hard. I could tell indoor from outdoor, basic guesses, but when it comes to the details—'Is it a restaurant?'—those details are missing. Even for annotation purposes, the task is really hard."

Complementary modalities

Because it takes far less power to collect and process audio data than it does to collect and process visual data, the researchers envision that a sound-recognition system could be used to improve the context sensitivity of mobile devices.

When coupled with GPS data, for instance, a sound-recognition system could determine that a cellphone user is in a movie theater and that the movie has started, and the phone could automatically route calls to a prerecorded outgoing message. Similarly, sound recognition could improve the situational awareness of autonomous robots.

"For instance, think of a self-driving car," Aytar says. "There's an ambulance coming, and the car doesn't see it. If it hears it, it can make future predictions for the ambulance—which path it's going to take—just purely based on sound."

Visual language

The researchers' machine-learning system is a neural network, so called because its architecture loosely resembles that of the human brain. A neural net consists of processing nodes that, like individual neurons, can perform only rudimentary computations but are densely interconnected. Information—say, the pixel values of a digital image—is fed to the bottom layer of nodes, which processes it and feeds it to the next layer, which processes it and feeds it to the next layer, and so on. The training process continually modifies the settings of the individual nodes, until the output of the final layer reliably performs some classification of the data—say, identifying the objects in the image.

Vondrick, Aytar, and Torralba first trained a neural net on two large, annotated sets of images: one, the ImageNet data set, contains labeled examples of images of 1,000 different objects; the other, the Places data set created by Torralba's group, contains labeled images of 401 different scene types, such as a playground, bedroom, or conference room.

Once the network was trained, the researchers fed it the video from 26 terabytes of video data downloaded from the photo-sharing site Flickr. "It's about 2 million unique videos," Vondrick says. "If you were to watch all of them back to back, it would take you about two years." Then they trained a second neural network on the audio from the same videos. The second network's goal was to correctly predict the object and scene tags produced by the first network.

The result was a network that could interpret natural sounds in terms of image categories. For instance, it might determine that the sound of birdsong tends to be associated with forest scenes and pictures of trees, birds, birdhouses, and bird feeders.

Benchmarking

To compare the sound-recognition network's performance to that of its predecessors, however, the researchers needed a way to translate its language of images into the familiar language of sound names. So they trained a simple machine-learning system to associate the outputs of the sound-recognition network with a set of standard sound labels.

For that, the researchers did use a database of annotated audio—one with 50 categories of sound and about 2,000 examples. Those annotations had been supplied by humans. But it's much easier to label 2,000 examples than to label 2 million. And the MIT researchers' network, trained first on unlabeled video, significantly outperformed all previous networks trained solely on the 2,000 labeled examples.

More information: SoundNet: Learning sound representations from unlabeled video, web.mit.edu/vondrick/soundnet.pdf

Provided by Massachusetts Institute of Technology

Citation: Computer learns to recognize sounds by watching video (2016, December 1) retrieved 26 April 2024 from https://phys.org/news/2016-12-video.html