TechXplore

System correlates recorded speech with images, could lead to fully automated speech recognition

December 7 2016, by Larry Hardesty

MIT researchers have developed a new approach to training speech-recognition systems that doesn't depend on transcription. Instead, their system analyzes correspondences between images and spoken descriptions of those images, as captured in a large collection of audio recordings. Credit: Massachusetts Institute of Technology

Speech recognition systems, such as those that convert speech to text on cellphones, are generally the result of machine learning. A computer pores through thousands or even millions of audio files and their transcriptions, and learns which acoustic features correspond to which typed words.

But transcribing recordings is costly, time-consuming work, which has limited speech recognition to a small subset of languages spoken in wealthy nations.

At the Neural Information Processing Systems conference this week, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) are presenting a new approach to training speech-recognition systems that doesn't depend on transcription. Instead, their system analyzes correspondences between images and spoken descriptions of those images, as captured in a large collection of audio recordings. The system then learns which acoustic features of the recordings correlate with which image characteristics.

"The goal of this work is to try to get the machine to learn language more like the way humans do," says Jim Glass, a senior research scientist at CSAIL and a co-author on the paper describing the new system. "The current methods that people use to train up speech recognizers are very supervised. You get an utterance, and you're told what's said. And you do this for a large body of data.

"Big advances have been made—Siri, Google—but it's expensive to get those annotations, and people have thus focused on, really, the major languages of the world. There are 7,000 languages, and I think less than 2 percent have ASR [automatic speech recognition] capability, and probably nothing is going to be done to address the others. So if you're trying to think about how technology can be beneficial for society at large, it's interesting to think about what we need to do to change the current situation. And the approach we've been taking through the years is looking at what we can learn with less supervision."

Joining Glass on the paper are first author David Harwath, a graduate student in electrical engineering and computer science (EECS) at MIT; and Antonio Torralba, an EECS professor.

Visual semantics

The version of the system reported in the new paper doesn't correlate recorded speech with written text; instead, it correlates speech with groups of thematically related images. But that correlation could serve as the basis for others.

If, for instance, an utterance is associated with a particular class of images, and the images have text terms associated with them, it should be possible to find a likely transcription of the utterance, all without human intervention. Similarly, a class of images with associated text terms in different languages could provide a way to do automatic translation.

Conversely, text terms associated with similar clusters of images, such as, say, "storm" and "clouds," could be inferred to have related meanings. Because the system in some sense learns words' meanings—the images associated with them—and not just their sounds, it has a wider range of potential applications than a standard speech recognition system.

To test their system, the researchers used a database of 1,000 images, each of which had a recording of a free-form verbal description associated with it. They would feed their system one of the recordings and ask it to retrieve the 10 images that best matched it. That set of 10 images would contain the correct one 31 percent of the time.

"I always emphasize that we're just taking baby steps here and have a long way to go," Glass says. "But it's an encouraging start."

The researchers trained their system on images from a huge database built by Torralba; Aude Oliva, a principal research scientist at CSAIL; and their students. Through Amazon's Mechanical Turk crowdsourcing site, they hired people to describe the images verbally, using whatever phrasing came to mind, for about 10 to 20 seconds.

For an initial demonstration of the researchers' approach, that kind of tailored data was necessary to ensure good results. But the ultimate aim is to train the system using digital video, with minimal human involvement. "I think this will extrapolate naturally to video," Glass says.

Merging modalities

To build their system, the researchers used neural networks, machine-learning systems that approximately mimic the structure of the brain. Neural networks are composed of processing nodes that, like individual neurons, are capable of only very simple computations but are connected to each other in dense networks. Data is fed to a network's input nodes, which modify it and feed it to other nodes, which modify it and feed it to still other nodes, and so on. When a neural network is being trained, it constantly modifies the operations executed by its nodes in order to improve its performance on a specified task.

The researchers' network is, in effect, two separate networks: one that takes images as input and one that takes spectrograms, which represent audio signals as changes of amplitude, over time, in their component frequencies. The output of the top layer of each network is a 1,024-dimensional vector—a sequence of 1,024 numbers.

The final node in the network takes the dot product of the two vectors.

That is, it multiplies the corresponding terms in the vectors together and adds them all up to produce a single number. During training, the networks had to try to maximize the dot product when the audio signal corresponded to an image and minimize it when it didn't.

For every spectrogram that the researchers' system analyzes, it can identify the points at which the dot-product peaks. In experiments, those peaks reliably picked out words that provided accurate image labels—"baseball," for instance, in a photo of a baseball pitcher in action, or "grassy" and "field" for an image of a grassy field.

In ongoing work, the researchers have refined the system so that it can pick out spectrograms of individual words and identify just those regions of an image that correspond to them.

More information: Paper: "Unsupervised learning of spoken language with visual context" papers.nips.cc/paper/6186-unsu … h-visual-context.pdf

Provided by Massachusetts Institute of Technology

Citation: System correlates recorded speech with images, could lead to fully automated speech recognition (2016, December 7) retrieved 5 May 2024 from https://techxplore.com/news/2016-12-speech-images-fully-automated-recognition.html