Machine-learning system tackles speech and object recognition, all at once

18 September 2018, by Rob Matheson

MIT computer scientists have developed a system that learns to identify objects within an image, based on a spoken description of the image. Credit: Christine Daniloff

MIT computer scientists have developed a system that learns to identify objects within an image, based on a spoken description of the image. Given an image and an audio caption, the model will highlight in real-time the relevant regions of the image being described.

Unlike current speech-recognition technologies, the model doesn't require manual transcriptions and annotations of the examples it's trained on. Instead, it learns words directly from recorded speech clips and objects in raw images, and associates them with one another.

The model can currently recognize only several hundred different words and object types. But the researchers hope that one day their combined speech-object recognition technique could save countless hours of manual labor and open new doors in speech and image recognition.

Speech-recognition systems such as Siri and Google Voice, for instance, require transcriptions of many thousands of hours of speech recordings. Using these data, the systems learn to map speech signals with specific words. Such an approach becomes especially problematic when, say, new terms enter our lexicon, and the systems must be retrained.

"We wanted to do speech recognition in a way that's more natural, leveraging additional signals and information that humans have the benefit of using, but that machine learning algorithms don't typically have access to. We got the idea of training a model in a manner similar to walking a child through the world and narrating what you're seeing," says David Harwath, a researcher in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and the Spoken Language Systems Group. Harwath co-authored a paper describing the model that was presented at the recent European Conference on Computer Vision.

In the paper, the researchers demonstrate their model on an image of a young girl with blonde hair and blue eyes, wearing a blue dress, with a white lighthouse with a red roof in the background. The model learned to associate which pixels in the image corresponded with the words "girl," "blonde hair," "blue eyes," "blue dress," "white light house," and "red roof." When an audio caption was narrated, the model then highlighted each of those objects in the image as they were described.

One promising application is learning translations between different languages, without need of a bilingual annotator. Of the estimated 7,000 languages spoken worldwide, only 100 or so have enough transcription data for speech recognition. Consider, however, a situation where two different-language speakers describe the same image. If the model learns speech signals from language A that correspond to objects in the image, and learns the signals in language B that correspond to those same objects, it could assume those two signals—and matching words—are translations of one another.

"There's potential there for a Babel Fish-type of mechanism," Harwath says, referring to the fictitious living earpiece in the "Hitchhiker's Guide to the Galaxy" novels that translates different languages to the wearer.

The CSAIL co-authors are: graduate student Adria Recasens; visiting student Didac Suris; former researcher Galen Chuang; Antonio Torralba, a professor of electrical engineering and computer science who also heads the MIT-IBM Watson AI Lab; and Senior Research Scientist James Glass, who leads the Spoken Language Systems Group at CSAIL.

Audio-visual associations

This work expands on an earlier model developed by Harwath, Glass, and Torralba that correlates speech with groups of thematically related images. In the earlier research, they put images of scenes from a classification database on the crowdsourcing Mechanical Turk platform. They then had people describe the images as if they were narrating to a child, for about 10 seconds. They compiled more than 200,000 pairs of images and audio captions, in hundreds of different categories, such as beaches, shopping malls, city streets, and bedrooms.

They then designed a model consisting of two separate convolutional neural networks (CNNs). One processes images, and one processes spectrograms, a visual representation of audio signals as they vary over time. The highest layer of the model computes outputs of the two networks and maps the speech patterns with image data.

The researchers would, for instance, feed the model caption A and image A, which is correct. Then, they would feed it a random caption B with image A, which is an incorrect pairing. After comparing thousands of wrong captions with image A, the model learns the speech signals corresponding with image A, and associates those signals with words in the captions. As described in a 2016 study, the model learned, for instance, to pick out the signal corresponding to the word "water," and to retrieve images with bodies of water.

"But it didn't provide a way to say, 'This is exact point in time that somebody said a specific word that refers to that specific patch of pixels,'" Harwath says.

Making a matchmap

In the new paper, the researchers modified the model to associate specific words with specific patches of pixels. The researchers trained the model on the same database, but with a new total of 400,000 image-captions pairs. They held out 1,000 random pairs for testing.

In training, the model is similarly given correct and incorrect images and captions. But this time, the image-analyzing CNN divides the image into a grid of cells consisting of patches of pixels. The audio-analyzing CNN divides the spectrogram into segments of, say, one second to capture a word or two.

With the correct image and caption pair, the model matches the first cell of the grid to the first segment of audio, then matches that same cell with the second segment of audio, and so on, all the way through each grid cell and across all time segments. For each cell and audio segment, it provides a similarity score, depending on how closely the signal corresponds to the object.

The challenge is that, during training, the model doesn't have access to any true alignment information between the speech and the image. "The biggest contribution of the paper," Harwath says, "is demonstrating that these cross-modal [audio and visual] alignments can be inferred automatically by simply teaching the network which images and captions belong together and which pairs don't."

The authors dub this automatic-learning association between a spoken caption's waveform with the image pixels a "matchmap." After training on thousands of image-caption pairs, the network narrows down those alignments to specific words representing specific objects in that matchmap.

"It's kind of like the Big Bang, where matter was really dispersed, but then coalesced into planets and stars," Harwath says. "Predictions start dispersed everywhere but, as you go through training, they converge into an alignment that represents meaningful semantic groundings between spoken words and visual objects."

More information: Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input. arxiv.org/abs/1804.01452

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
APA citation: Machine-learning system tackles speech and object recognition, all at once (2018, September 18) retrieved 4 December 2020 from https://techxplore.com/news/2018-09-machine-learning-tackles-speech-recognition.html