DAILY NEWS 3 November 2016

Binge-watching videos teaches computers to recognise sounds

Future security systems could listen out for break-ins
Blend Images/Superstock

By **Aviva Rutkin**

Now machines are going on internet-watching sprees too – but with something to show for it. After viewing a year's worth of online videos, a computer model has learned to distinguish between sounds such as bird chirps, door knocks, snoring and fireworks.

Such technology could transform how we interact with machines and make it easier for our cellphones, smart homes and robot assistants to understand the world around them.

Computer vision has dramatically improved over the past few years thanks to the wealth of labelled data machines can tap into online. They can now recognise faces or cats as accurately as a human can.

But their listening abilities still lag behind because there is not nearly as much useful sound data available.

One group of computer scientists wondered if they could piggyback on the advances made in computer vision to improve machine listening.

Sound and vision

"We thought: 'We can actually transfer this visual knowledge that's been learned by machines to another domain where we don't have any data, but we do have this natural synchronisation between images and sounds,'" says Yusuf Aytar at the Massachusetts Institute of Technology.

Aytar and his colleagues Carl Vondrick and Antonio Torralba downloaded over two million videos from Flickr, representing a total running time of more than a year. The computer effectively marathoned through the videos, first picking out the objects in the shot, then comparing what it saw to the raw sound.

If it picked up on the visual features of babies in different videos, for example, and found they often appeared alongside babbling noises, it learned to identify that sound as a baby's babble even without the visual clue.

"It's learning from these videos without any human in the loop," says Vondrick. "It's learning in some sense on its own to recognise sound from just a year of video."

The researchers tested several versions of their SoundNet model on three data sets, asking it to sort between sounds such as rain, sneezes, ticking clocks and roosters. At its best, the computer was 92.2 per cent accurate. Humans scored 95.7 per cent on the same challenge.

Laughing hens?

A few sounds still give the SoundNet trouble, however. It might mistake footsteps for door knocks, for instance, or

insects for washing machines. It sometimes also confuses laughter with the sound of hens. But more training could help it sort out those fine details.

The study will be presented next month at the Neural Information Processing Systems conference in Barcelona, Spain.

"This is like nothing we've seen before," says Ian McLoughlin at the University of Kent in the UK.

Most of us communicate primarily using speech and hearing, so advances like this mean we could one day speak to machines in a much more natural way. "In human-computer interaction, up to today, we've really just explored vision," McLoughlin says. "We've used our eyes to look at graphics – that's what computers do. But the next dimension is audio."

For example, many of us struggle to get a voice-activated digital assistant such as Apple's Siri to understand what we are saying because it misses words or picks up on irrelevant noise.

Natural communication

With more listening smarts, these assistants could communicate more naturally with you and not be confused if your speech is interrupted by a distracting noise such as an ambulance siren or a dog barking. It could even use such background sounds to understand the context of a situation.

"Microphones are much cheaper and use much less power than a camera," says Vondrick. "If you want to deploy this on your phone, it wouldn't drain your battery as much as if you had your camera on all the time."

Home security could be another valuable application. Companies such as Audio Analytic in Cambridge, UK, aim to help people protect their properties by listening for threatening sounds – like a window shattering or a smoke alarm blaring. Programs like SoundNet make that goal more feasible.

"This would allow you to set up a security system or perhaps interrogate your smart home to find out what's happening in the home," says Mark Plumbley at the University of Surrey in the UK. "With recent announcements from Google and Amazon of the Google Home assistant and the Amazon Echo, the idea that a microphone might be around the home and on all the time now is something that could become quite common."

Journal reference: *arXiv*, DOI: 1610.09001

Read more: Auto twitcher recognises different bird songs even when noisy

A shorter version of this article was published in *New Scientist* magazine on 12 November 2016