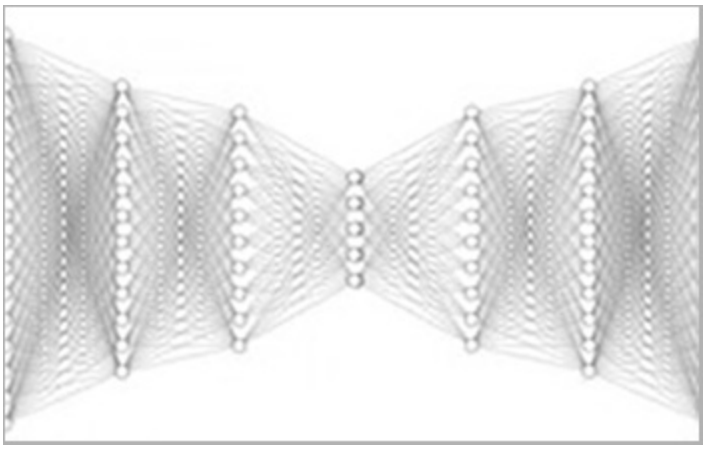
Research News

Deep learning networks may prefer the human voice -- as we do

AI systems might reach higher performance if programmed with human language



A deep neural network that is taught to speak demonstrates higher learning.

Credit and Larger Version (/discoveries/disc_images.jsp?cntn_id=302548&org=NSF)

April 22, 2021

The digital revolution is built on a foundation of binaries, invisible 1s and 0s called bits. The notion that computers prefer to "speak" in binary numbers is rarely questioned. According to new research from Columbia Engineering (/cgi-bin/good-bye?https://www.engineering.columbia.edu/press-releases/hod-lipson-deep-learning-networks-prefer-human-voice), that could be about to change.

A new U.S. National Science Foundation <https://www.nsf.gov/awardsearch/showAward?AWD_ID=1925157&HistoricalAwards=false> -funded study by mechanical engineer Hod Lipson and researcher Boyuan Chen proves that artificial intelligence systems might reach higher levels of performance if they are programmed with sound files of human language rather than with numerical data labels.

The researchers discovered that a neural network whose "training labels" consisted of sound files reached higher levels of performance in identifying objects in images than another network that had been programmed in a more traditional manner that used simple binary inputs.

"To understand why this finding is significant," said Lipson, "it's useful to understand how neural networks are usually programmed, and why using the sound of the human voice is a radical experiment."

The language of binary numbers conveys information compactly and precisely. In contrast, spoken human language is more tonal and analog, and, when captured in a digital file, non-binary. Because numbers are such an efficient way to digitize data, programmers rarely deviate from a numbers-driven process when they develop a neural network.

Lipson and Chen speculated that neural networks might learn faster and better if the systems were "trained" to recognize objects, animals, for instance, by using the power of one of the world's most highly evolved sounds -- the human voice uttering specific words.

The team set up the experimental neural network in a novel way. They fed it a data table containing a photograph of an animal or object and an audio file of human voicing of the word for the depicted animal or object. There were no 1s and 0s.

At first, the researchers were somewhat surprised to discover that their hunch had been correct -- there was no apparent advantage between the audio file and the binary 1s and 0s. Both the control neural network and the experimental one performed equally well, correctly identifying the animal or object depicted in a photograph about 92% of the time. To double-check their results, the researchers ran the experiment again and got the same outcome.

The results, to be presented at the International Conference on Learning Representations (/cgi-bin/good-bye?https://iclr.cc/) on May 3, is part of a broader effort at Lipson's Columbia Creative Machines Lab to create robots that can understand the world around them by interacting with other machines and humans rather than by being programmed directly with carefully preprocessed data.

-- NSF Public Affairs, researchnews@nsf.gov (mailto:researchnews@nsf.gov)
