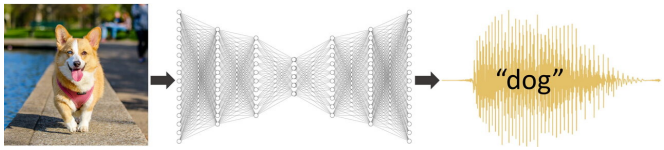
Deep learning networks prefer the human voice—just like us

6 April 2021



A deep neural network that is taught to speak out the answer demonstrates higher performances of learning robust and efficient features. This study opens up new research questions on the role of label representations for object recognition. Credit: Creative Machines Lab/Columbia Engineering

The digital revolution is built on a foundation of invisible 1s and 0s called bits. As decades pass, and more and more of the world's information and knowledge morph into streams of 1s and 0s, the notion that computers prefer to "speak" in binary numbers is rarely questioned. According to new research from Columbia Engineering, this could be about to change.

A new study from Mechanical Engineering Professor Hod Lipson and his Ph.D. student Boyuan Chen proves that artificial intelligence systems might actually 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 in a side-by-side comparison, a neural network whose "training labels" consisted of sound files reached higher levels of performance in identifying objects in images, compared to another network that had been programmed in a more traditional manner, using simple binary inputs.

"To understand why this finding is significant," said Lipson, James and Sally Scapa Professor of Innovation and a member of Columbia's Data Science Institute, "It's useful to understand how neural networks are usually programmed, and why using the sound of the human voice is a radical experiment."

When used to convey information, the language of binary numbers is compact and precise. 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, a highly regarded roboticist, and Chen, a former concert pianist, had a hunch that neural networks might not be reaching their full potential. They speculated that neural networks might learn faster and better if the systems were "trained" to recognize animals, for instance, by using the power of one of the world's most highly evolved sounds—the human voice uttering specific words.

One of the more common exercises AI researchers use to test out the merits of a new machine learning technique is to train a neural network to recognize specific objects and animals in a collection of different photographs. To check their hypothesis, Chen, Lipson and two students, Yu Li and Sunand Raghupathi, set up a controlled experiment. They created two new neural networks with the goal of training both of them to recognize 10 different types of objects in a collection of 50,000 photographs known as "training images."

One AI system was trained the traditional way, by uploading a giant data table containing thousands of rows, each row corresponding to a single training photo. The first column was an image file containing a photo of a particular object or animal; the next 10 columns corresponded to 10 possible object types: cats, dogs, airplanes, etc. A "1" in any column indicates the correct answer, and nine 0s indicate the incorrect answers.

The team set up the experimental neural network in

a radically novel way. They fed it a data table whose rows contained a photograph of an animal or object, and the second column contained an audio file of a recorded human voice actually voicing the word for the depicted animal or object out loud. There were no 1s and 0s.

Once both neural networks were ready, Chen, Li, and Raghupathi trained both AI systems for a total of 15 hours and then compared their respective performance. When presented with an image, the original network spat out the answer as a series of ten 1s and 0s—just as it was trained to do. The experimental neural network, however, produced a clearly discernible voice trying to "say" what the object in the image was. Initially the sound was just a garble. Sometimes it was a confusion of multiple categories, like "cog" for cat and dog. Eventually, the voice was mostly correct, albeit with an eerie alien tone (see example on website).

At first, the researchers were somewhat surprised to discover that their hunch had been correct—there was no apparent advantage to 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.

What they discovered next, however, was even more surprising. To further explore the limits of using sound as a training tool, the researchers set up another side-by-side comparison, this time using far fewer photographs during the training process. While the first round of training involved feeding both neural networks data tables containing 50,000 training images, both systems in the second experiment were fed far fewer training photographs, just 2,500 apiece.

It is well known in AI research that most neural networks perform poorly when training data is sparse, and in this experiment, the traditional, numerically trained network was no exception. Its ability to identify individual animals that appeared in the photographs plummeted to about 35% accuracy. In contrast, although the experimental neural network was also trained with the same number of photographs, its performance did twice as well, dropping only to 70% accuracy.

Intrigued, Lipson and his students decided to test their voice-driven training method on another classic AI image recognition challenge, that of image ambiguity. This time they set up yet another side-by-side comparison but raised the game a notch by using more difficult photographs that were harder for an AI system to "understand." For example, one training photo depicted a slightly corrupted image of a dog, or a cat with odd colors. When they compared results, even with more challenging photographs, the voice-trained neural network was still correct about 50% of the time, outperforming the numerically-trained network that floundered, achieving only 20% accuracy.

Ironically, the fact their results went directly against the status quo became challenging when the researchers first tried to share their findings with their colleagues in computer science. "Our findings run directly counter to how many experts have been trained to think about computers and numbers; it's a common assumption that binary inputs are a more efficient way to convey information to a machine than audio streams of similar information 'richness,'" explained Boyuan Chen, the lead researcher on the study. "In fact, when we submitted this research to a big AI conference, one anonymous reviewer rejected our paper simply because they felt our results were just 'too surprising and un-intuitive.'"

When considered in the broader context of information theory however, Lipson and Chen's hypothesis actually supports a much older, landmark hypothesis first proposed by the legendary Claude Shannon, the father of information theory. According to Shannon's theory, the most effective communication "signals" are characterized by an optimal number of bits, paired with an optimal amount of useful information, or "surprise."

"If you think about the fact that human language has been going through an optimization process for tens of thousands of years, then it makes perfect sense, that our spoken words have found a good balance between noise and signal;" Lipson

observed. "Therefore, when viewed through the lens of Shannon Entropy, it makes sense that a neural network trained with human language would outperform a neural network trained by simple 1s and 0s."

The study, to be presented at the International Conference on Learning Representations conference on May 3, 2021, 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 programed directly with carefully preprocessed data.

"We should think about using novel and better ways to train AI systems instead of collecting larger datasets," said Chen. "If we rethink how we present training data to the machine, we could do a better job as teachers."

One of the more refreshing results of computer science research on artificial intelligence has been an unexpected side effect: by probing how machines learn, sometimes researchers stumble upon fresh insight into the grand challenges of other, well-established fields.

"One of the biggest mysteries of human evolution is how our ancestors acquired language, and how children learn to speak so effortlessly," Lipson said. "If human toddlers learn best with repetitive spoken instruction, then perhaps AI systems can, too."

More information: Project web site: www.creativemachineslab.com/la … -representation.html

Paper: openreview.net/pdf?id=MyHwDabUHZm

Provided by Columbia University School of Engineering and Applied Science