Can machine-learning models overcome biased datasets?

February 21 2022, by Adam Zewe

Credit: CC0 Public Domain

Artificial intelligence systems may be able to complete tasks quickly, but that doesn't mean they always do so fairly. If the datasets used to train machine-learning models contain biased data, it is likely the system could exhibit that same bias when it makes decisions in practice.

For instance, if a dataset contains mostly images of white men, then a facial-recognition model trained with this data may be less accurate for women or people with different skin tones.

A group of researchers at MIT, in collaboration with researchers at Harvard University and Fujitsu, Ltd., sought to understand when and how a machine-learning model is capable of overcoming this kind of dataset bias. They used an approach from neuroscience to study how training data affects whether an artificial neural network can learn to recognize objects it has not seen before. A neural network is a machine-learning model that mimics the human brain in the way it contains layers of interconnected nodes, or "neurons," that process data.

The new results show that diversity in training data has a major influence on whether a neural network is able to overcome bias, but at the same time dataset diversity can degrade the network's performance. They also show that how a neural network is trained, and the specific types of neurons that emerge during the training process, can play a major role in whether it is able to overcome a biased dataset.

"A neural network can overcome dataset bias, which is encouraging. But the main takeaway here is that we need to take into account data diversity. We need to stop thinking that if you just collect a ton of raw data, that is going to get you somewhere. We need to be very careful about how we design datasets in the first place," says Xavier Boix, a research scientist in the Department of Brain and Cognitive Sciences (BCS) and the Center for Brains, Minds, and Machines (CBMM), and senior author of the paper.

Co-authors include former graduate students Spandan Madan, a corresponding author who is currently pursuing a Ph.D. at Harvard, Timothy Henry, Jamell Dozier, Helen Ho, and Nishchal Bhandari; Tomotake Sasaki, a former visiting scientist now a researcher at Fujitsu; Frédo Durand, a professor of electrical engineering and computer science and a member of the Computer Science and Artificial Intelligence Laboratory; and Hanspeter Pfister, the An Wang Professor of Computer Science at the Harvard School of Enginering and Applied Sciences. The research appears today in *Nature Machine Intelligence*.

Thinking like a neuroscientist

Boix and his colleagues approached the problem of dataset bias by thinking like neuroscientists. In neuroscience, Boix explains, it is common to use controlled datasets in experiments, meaning a dataset in which the researchers know as much as possible about the information it contains.

The team built datasets that contained images of different objects in varied poses, and carefully controlled the combinations so some datasets had more diversity than others. In this case, a dataset had less diversity if it contains more images that show objects from only one viewpoint. A more diverse dataset had more images showing objects from multiple viewpoints. Each dataset contained the same number of images.

The researchers used these carefully constructed datasets to train a neural network for image classification, and then studied how well it was able to identify objects from viewpoints the network did not see during training (known as an out-of-distribution combination).

For example, if researchers are training a model to classify cars in images, they want the model to learn what different cars look like. But if every Ford Thunderbird in the training dataset is shown from the front, when the trained model is given an image of a Ford Thunderbird shot from the side, it may misclassify it, even if it was trained on millions of car photos.

The researchers found that if the dataset is more diverse—if more images show objects from different viewpoints—the network is better able to generalize to new images or viewpoints. Data diversity is key to overcoming bias, Boix says.

"But it is not like more data diversity is always better; there is a tension here. When the neural network gets better at recognizing new things it hasn't seen, then it will become harder for it to recognize things it has already seen," he says.

Testing training methods

The researchers also studied methods for training the neural network.

In machine learning, it is common to train a network to perform multiple tasks at the same time. The idea is that if a relationship exists between the tasks, the network will learn to perform each one better if it learns them together.

But the researchers found the opposite to be true—a model trained separately for each task was able to overcome bias far better than a model trained for both tasks together.

"The results were really striking. In fact, the first time we did this experiment, we thought it was a bug. It took us several weeks to realize it was a real result because it was so unexpected," he says.

They dove deeper inside the neural networks to understand why this occurs.

They found that neuron specialization seems to play a major role. When the neural network is trained to recognize objects in images, it appears that two types of neurons emerge—one that specializes in recognizing the object category and another that specializes in recognizing the viewpoint.

When the network is trained to perform tasks separately, those specialized neurons are more prominent, Boix explains. But if a network is trained to do both tasks simultaneously, some neurons become diluted and don't specialize for one task. These unspecialized neurons are more likely to get confused, he says.

"But the next question now is, how did these neurons get there? You train the neural network and they emerge from the learning process. No one told the network to include these types of neurons in its architecture. That is the fascinating thing," he says.

That is one area the researchers hope to explore with future work. They want to see if they can force a neural network to develop neurons with this specialization. They also want to apply their approach to more complex tasks, such as objects with complicated textures or varied illuminations.

Boix is encouraged that a neural network can learn to overcome bias, and he is hopeful their work can inspire others to be more thoughtful about the datasets they are using in AI applications.

More information: Spandan Madan, When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations, *Nature Machine Intelligence* (2022). DOI: 10.1038/s42256-021-00437-5.
www.nature.com/articles/s42256-021-00437-5

Provided by Massachusetts Institute of Technology

Citation: Can machine-learning models overcome biased datasets? (2022, February 21) retrieved 27 April 2024 from https://techxplore.com/news/2022-02-machine-learning-biased-datasets.html