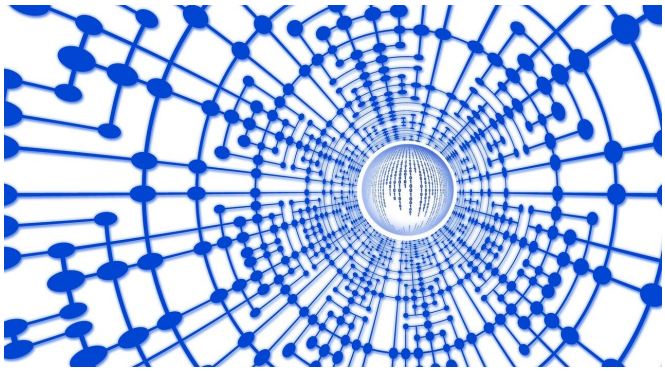
A math idea that may dramatically reduce the dataset size needed to train AI systems

23 October 2020, by Bob Yirka



Credit: CC0 Public Domain

A pair of statisticians at the University of Waterloo has proposed a math process idea that might allow for teaching AI systems without the need for a large dataset. Ilia Sucholutsky and Matthias Schonlau have written a paper describing their idea and published it on the arXiv preprint server.

Artificial intelligence (AI) applications have been the subject of much research lately, with the development of deep learning networks, researchers in a wide range of fields began finding uses for it, including creating deepfake videos, board game applications and medical diagnostics.

Deep learning networks require large datasets in order to detect patterns revealing how to perform a given task, such as picking a certain face out of a crowd. In this new effort, the researchers wondered if there might be a way to reduce the size of the dataset. They noted that children only need to see a couple of pictures of an animal to recognize other examples. Being statisticians, they wondered if there might be a way to use mathematics to solve the problem.

The researchers built on recent work by a team at MIT. They had found that distilling the most pertinent information describing handwritten numbers in a dataset known as MNIST and packing them together greatly reduced the number of characters their AI system needed to learn to recognize letters in a new dataset. The pair in Canada noted that the reason the system was able to learn with much less data was because it was trained to recognize numbers in a new way: instead of just showing it the number 3 thousands of times, they trained it to recognize that the target was a number that looked somewhat (30 percent) like the digit 8, and so on with other digits. They called these hints soft labels.

They then took this idea further by applying it to a type of machine learning called k-nearest neighbor (kNN), which allowed them to transfer their idea into a graphical approach. And using that approach, they were able to apply soft labels to datasets describing XY coordinates on a graph. As a result, the AI system was easily trained to place dots on a graph on the correct side of a line they had drawn without the need for a large dataset. The researchers describe their approach as "less than one-shot learning" (LO-shot) and suggest it might be possible to expand it to other areas, though they acknowledge there is still one major hurdle to overcome. The system still requires a large dataset to start the winnowing process.

More information: 'Less Than One'-Shot Learning: Learning N Classes From M arxiv.org/abs/2009.08449