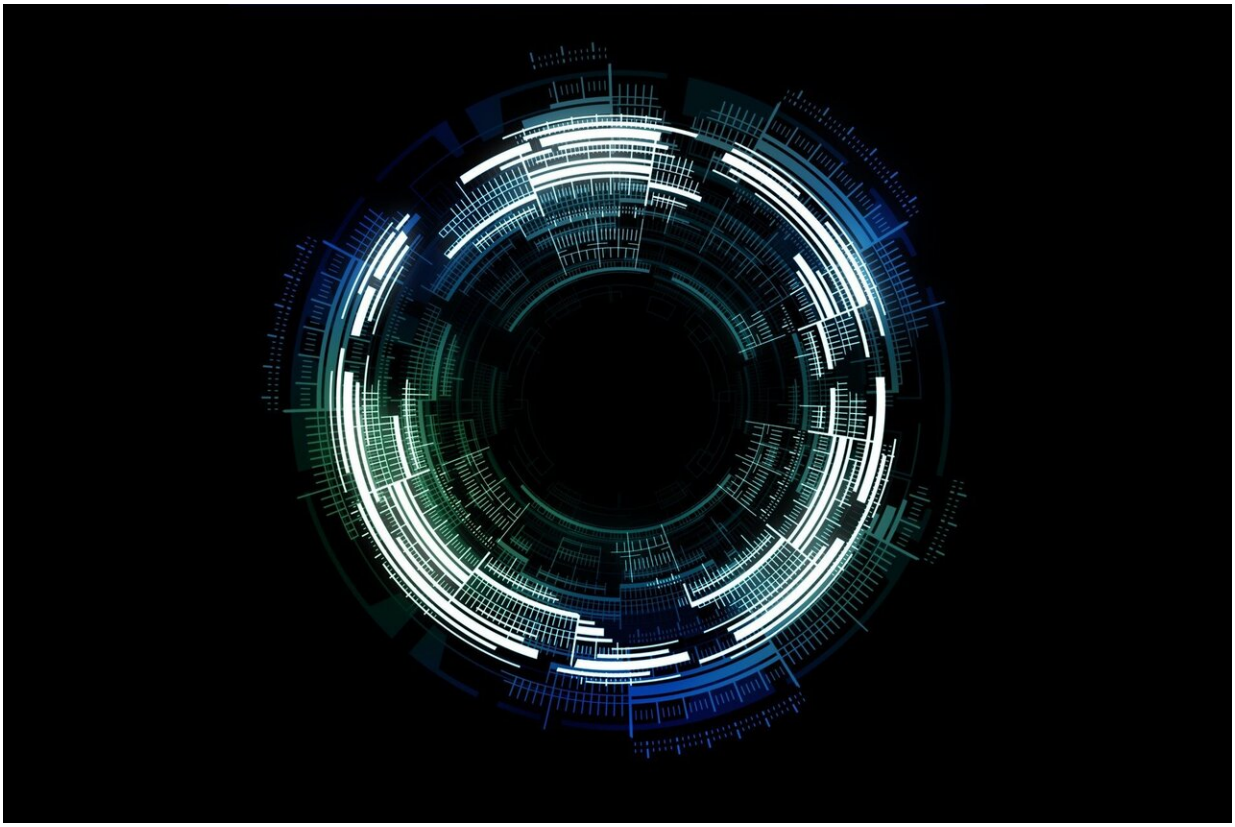
Machine learning applications need less data than has been assumed

July 28 2021, by Bob Yirka



Credit: Pixabay/CC0 Public Domain

A combined team of researchers from the University of British Columbia and the University of Alberta has found that at least some machine learning applications can learn from far fewer examples than

has been assumed. In their paper published in the journal *Nature Machine Intelligence*, the group describes testing they carried out with machine learning applications created to predict certain types of molecular structures.

Machine learning can be used in a wide variety of applications—one of the most well-known is learning to spot people or objects in photographs. Such applications typically require huge amounts of data for training. In this new effort, the researchers have found that in some instances, machine learning applications do not need such huge amounts of data to be useful.

The researchers were initially looking for ways to predict the structure of illegal designer drugs. Doing so would help medical researchers prepare for them should people consuming them begin showing up in hospital emergency rooms. The team realized their job would be much easier if they could use a machine learning application; unfortunately, there are only 1,700 known designer drugs that could be used to train such a system. Undaunted, the researchers wondered if it might be possible to figure out just how much data would be required for such a system to be useful, or if there might be a way to modify an algorithm or the data that was used to train it to allow for less available data.

To find out, the researchers created 8,500 models and trained each of them on differently sized datasets taken from the 500,000 molecules in the simplified molecular-input line-entry system. Then they used the models to predict possible molecular types. In so doing, they found many of the models worked quite well with the limited dataset. They also found that most of them began to level off in their predictive abilities after just 10,000 to 20,000 data records. When they used the best-performing models to conduct their initial research, they found the results were correct approximately 50% of the time.

More information: Michael A. Skinnider et al, Chemical language models enable navigation in sparsely populated chemical space, *Nature Machine Intelligence* (2021). DOI: 10.1038/s42256-021-00368-1