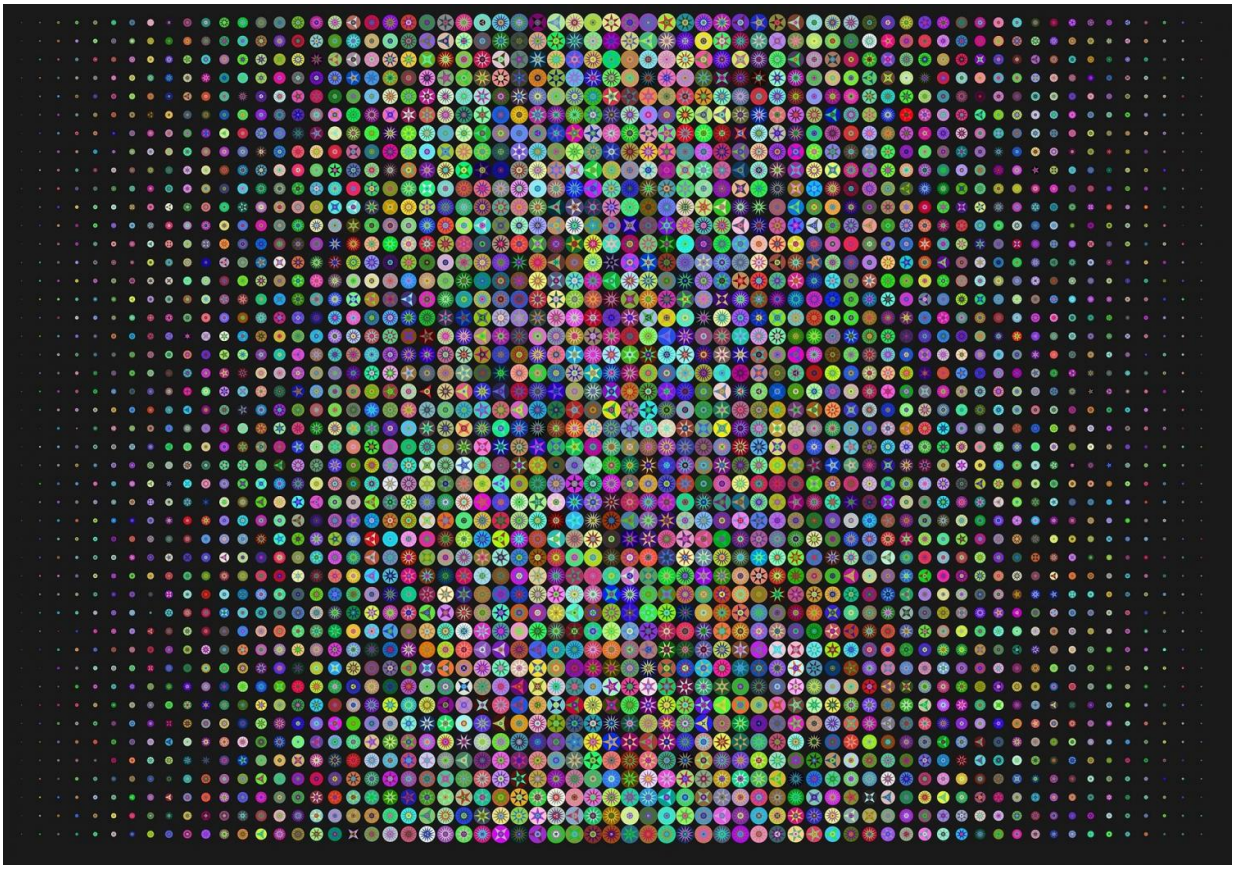
Researchers develop new algorithm that could reduce complexity of big data

March 9 2021, by Stephanie Jones



Credit: CC0 Public Domain

Whenever a scientific experiment is conducted, the results are turned into numbers, often producing huge datasets. In order to reduce the size of the data, computer programmers use algorithms that can find and

extract the principal features that represent the most salient statistical properties. But many such algorithms cannot be applied directly to these large volumes of data.

Reza Oftadeh, doctoral student in the Department of Computer Science and Engineering at Texas A&M University, advised by Dr. Dylan Shell, faculty in the department, developed an algorithm applicable to large datasets. It is a useful machine-learning tool because it can extract and directly order features from most salient to least.

"There are many ad hoc ways to extract these features using machine-learning algorithms, but we now have a fully rigorous theoretical proof that our model can find and extract these prominent features from the data simultaneously, doing so in one pass of the algorithm," said Oftadeh.

Their paper describing the research was published in the proceedings from the 2020 International Conference on Machine Learning.

A subfield of machine learning deals with component analysis, the problem of identifying and extracting a raw dataset's features to help reduce its dimensionality. Once identified, the features are used to make annotated samples of the data for further analysis or other machine-learning tasks such as classification, clustering, visualization and modeling based on those features.

The work to find or develop these types of algorithms has been going on for the past century, but what sets this era apart from the others is the existence of big data, which can contain many millions of sample points with 10s of thousands of attributes. Analyzing these massive datasets is a very complicated, time-consuming process for human programmers, so artificial neural networks (ANNs) have come to the forefront in recent years.

As one of the main tools of machine learning, ANNs are computational models that are designed to simulate how the human brain analyzes and processes information. They are typically made of dozens to millions of artificial neurons, called units, arranged in a series of layers that it uses to make sense of the information it's given. ANNs can be used in various ways, but they are most commonly used to identify the unique features that best represent the data and classify them into different categories based on that information.

"There are many ANNs that work very well, and we use them every day on our phones and computers," said Oftadeh. "For example, applications like Alexa, Siri and Google Translate utilize ANNs that are trained to recognize what different speech patterns, accents and voices are saying."

But not all features are equally significant, and they can be placed in order from most to least important. Previous approaches use a specific type of ANN called an autoencoder to extract them, but they cannot tell exactly where the features are located or which are more important than the others.

"For example, if you have hundreds of thousands of dimensions and want to find only 1,000 of the most prominent and order those 1,000, it is theoretically possible to do but not feasible in practice because the model would have to be run repeatedly on the dataset 1,000 times," said Oftadeh.

To make a more intelligent algorithm, the researchers propose adding a new cost function to the network that provides the exact location of the features directly ordered by their relative importance. Once incorporated, their method results in a more efficient processing that can be fed bigger datasets to perform classic data analysis.

To verify the effectiveness of their method, they trained their model for

an optical character recognition (OCR) experiment, which is the conversion of images of typed or handwritten text into machine-encoded text from inside digital physical documents, like a scanner produces. Once it's trained for OCR using the proposed method, the model can tell which features are most important.

Currently, the algorithm can only be applied to one-dimensional data samples, but the team is interested in extending their algorithm's abilities to handle even more complex structured data.

"Breaking down multidimensional data directly is a very active, challenging mathematical field of research with many challenges of its own, and we are interested in exploring it further," said Oftadeh.

The next step of their work is to generalize their method in a way that provides a unified framework to produce other machine-learning methods that can find the underlaying structure of a dataset and/or extract its features by setting a small number of specifications.

Other contributors to this research include Jiayi Shen, doctoral student in the computer science and engineering department, and Dr. Zhangyang "Atlas" Wang, assistant professor in the electrical and computer engineering department at The University of Texas at Austin. Also instrumental in identifying the research problem, and guiding Oftadeh, was Dr. Boris Hanin, assistant professor in the department of mathematics at Princeton University.

This research was funded by the National Science Foundation and U.S. Army Research Office Young Investigator Award.

More information: 2020 International Conference on Machine Learning Proceedings, proceedings.mlr.press/v119/oft … eh20a/oftadeh20a.pdf

Provided by Texas A&M University College of Engineering

Citation: Researchers develop new algorithm that could reduce complexity of big data (2021, March 9) retrieved 1 May 2024 from https://techxplore.com/news/2021-03-algorithm-complexity-big.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.