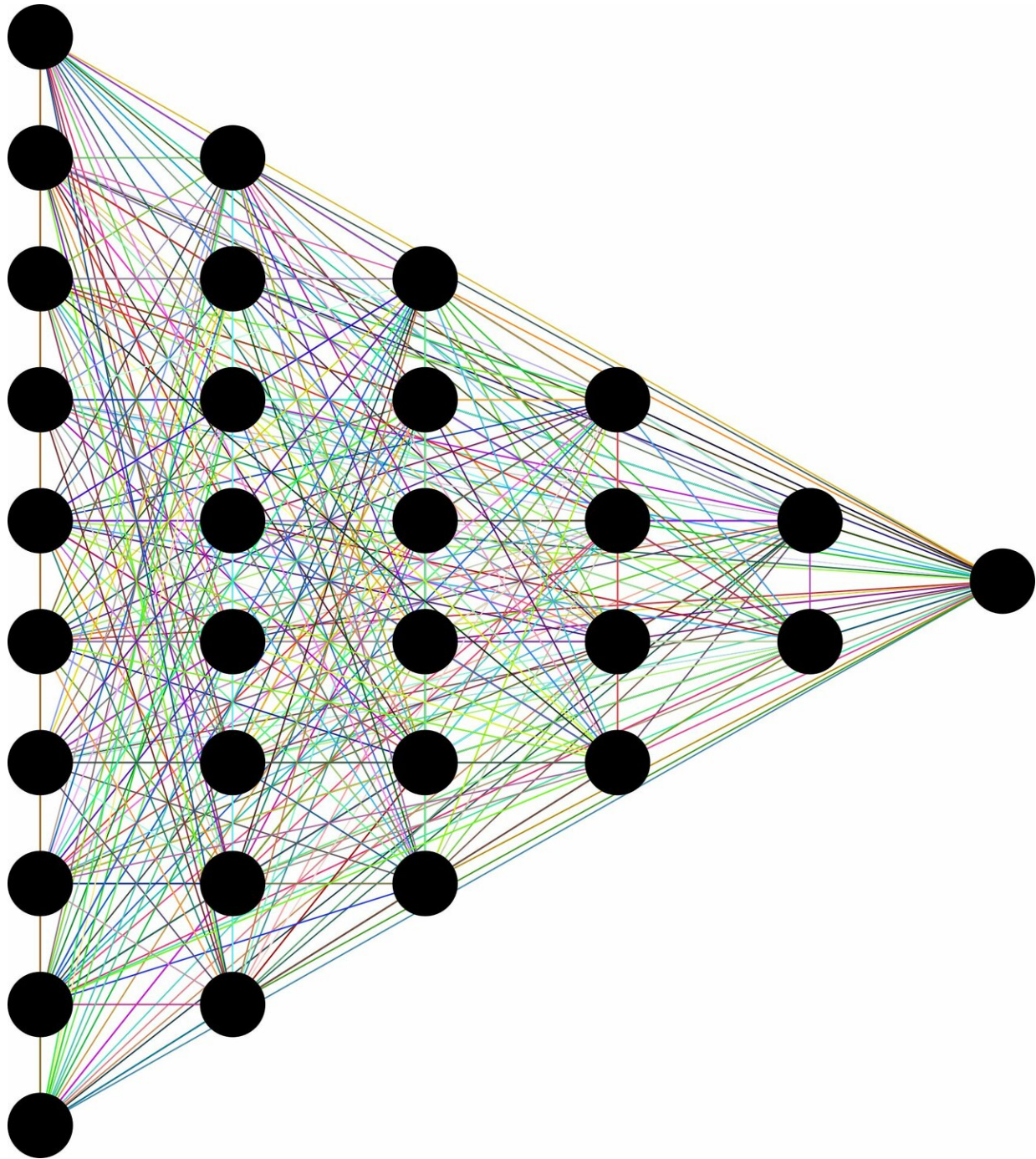


Information theory holds surprises for machine learning

January 24 2019



Credit: CC0 Public Domain

New SFI research challenges a popular conception of how machine

learning algorithms "think" about certain tasks.

The conception goes something like this: because of their ability to discard useless [information](#), a class of [machine learning](#) algorithms called [deep neural networks](#) can learn general concepts from raw data—like identifying cats generally after encountering tens of thousands of images of different cats in different situations. This seemingly human ability is said to arise as a byproduct of the networks' layered architecture. Early layers encode the "cat" label along with all of the raw information needed for [prediction](#). Subsequent layers then compress the information, as if through a bottleneck. Irrelevant data, like the color of the cat's coat, or the saucer of milk beside it, is forgotten, leaving only general features behind. Information theory provides bounds on just how optimal each layer is, in terms of how well it can balance the competing demands of [compression](#) and prediction.

"A lot of times when you have a neural network and it learns to map faces to names, or pictures to numerical digits, or amazing things like French text to English text, it has a lot of intermediate hidden layers that information flows through," says Artemy Kolchinsky, an SFI Postdoctoral Fellow and the study's lead author. "So there's this long-standing idea that as raw inputs get transformed to these intermediate representations, the system is trading prediction for compression, and building higher-level concepts through this information bottleneck."

However, Kolchinsky and his collaborators Brendan Tracey (SFI, MIT) and Steven Van Kuyk (University of Wellington) uncovered a surprising weakness when they applied this explanation to common classification problems, where each input has one correct output (e.g., in which each picture can either be of a cat or of a dog). In such cases, they found that classifiers with many layers generally do not give up some prediction for improved compression. They also found that there are many "trivial" representations of the inputs which are, from the point of view of

[information theory](#), optimal in terms of their balance between prediction and compression.

"We found that this information bottleneck measure doesn't see compression in the same way you or I would. Given the choice, it is just as happy to lump 'martini glasses' in with 'Labradors', as it is to lump them in with 'champagne flutes,'" Tracey explains. "This means we should keep searching for compression measures that better match our notions of compression."

While the idea of compressing inputs may still play a useful role in machine learning, this research suggests it is not sufficient for evaluating the internal representations used by different machine learning algorithms.

At the same time, Kolchinsky says that the concept of trade-off between compression and prediction will still hold for less deterministic tasks, like predicting the weather from a noisy dataset. "We're not saying that information bottleneck is useless for supervised [machine] learning," Kolchinsky stresses. "What we're showing here is that it behaves counter-intuitively on many common machine learning problems, and that's something people in the machine learning community should be aware of."

More information: Caveats for information bottleneck in deterministic scenarios. export.arxiv.org/abs/1808.07593

Provided by Santa Fe Institute

Citation: Information theory holds surprises for machine learning (2019, January 24) retrieved 25 April 2024 from <https://phys.org/news/2019-01-theory-machine.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.