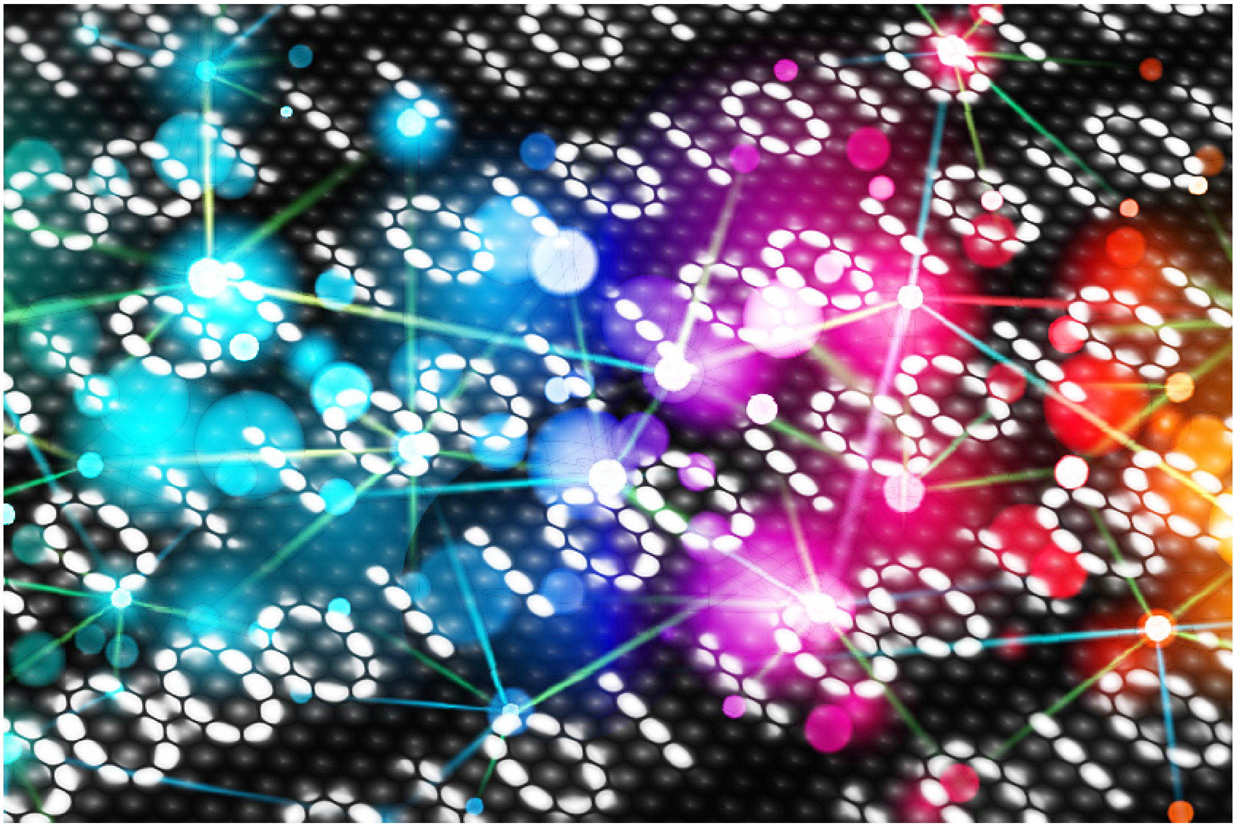
TechXplore

Technique reveals the basis for machine-learning systems' decisions

October 28 2016, by Larry Hardesty



Researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have devised a way to train neural networks so that they provide not only predictions and classifications but rationales for their decisions. Credit: Christine Daniloff/MIT

In recent years, the best-performing systems in artificial-intelligence research have come courtesy of neural networks, which look for patterns in training data that yield useful predictions or classifications. A neural net might, for instance, be trained to recognize certain objects in digital images or to infer the topics of texts.

But neural nets are black boxes. After training, a network may be very good at classifying data, but even its creators will have no idea why. With visual data, it's sometimes possible to automate experiments that determine which visual features a neural net is responding to. But text-processing systems tend to be more opaque.

At the Association for Computational Linguistics' Conference on Empirical Methods in Natural Language Processing, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) will present a new way to train neural networks so that they provide not only predictions and classifications but rationales for their decisions.

"In real-world applications, sometimes people really want to know why the model makes the predictions it does," says Tao Lei, an MIT graduate student in electrical engineering and computer science and first author on the new paper. "One major reason that doctors don't trust machine-learning methods is that there's no evidence."

"It's not only the medical domain," adds Regina Barzilay, the Delta Electronics Professor of Electrical Engineering and Computer Science and Lei's thesis advisor. "It's in any domain where the cost of making the wrong prediction is very high. You need to justify why you did it."

"There's a broader aspect to this work, as well," says Tommi Jaakkola, an MIT professor of electrical engineering and computer science and the third coauthor on the paper. "You may not want to just verify that the model is making the prediction in the right way; you might also want to exert some influence in terms of the types of predictions that it should make. How does a layperson communicate with a complex model that's trained with algorithms that they know nothing about? They might be able to tell you about the rationale for a particular prediction. In that sense it opens up a different way of communicating with the model."

Virtual brains

Neural networks are so called because they mimic—approximately—the structure of the brain. They are composed of a large number of processing nodes that, like individual neurons, are capable of only very simple computations but are connected to each other in dense networks.

In a process referred to as "deep learning," training data is fed to a network's input nodes, which modify it and feed it to other nodes, which modify it and feed it to still other nodes, and so on. The values stored in the network's output nodes are then correlated with the classification category that the network is trying to learn—such as the objects in an image, or the topic of an essay.

Over the course of the network's training, the operations performed by the individual nodes are continuously modified to yield consistently good results across the whole set of training examples. By the end of the process, the computer scientists who programmed the network often have no idea what the nodes' settings are. Even if they do, it can be very hard to translate that low-level information back into an intelligible description of the system's decision-making process.

In the new paper, Lei, Barzilay, and Jaakkola specifically address neural nets trained on textual data. To enable interpretation of a neural net's decisions, the CSAIL researchers divide the net into two modules. The first module extracts segments of text from the training data, and the segments are scored according to their length and their coherence: The shorter the segment, and the more of it that is drawn from strings of consecutive words, the higher its score.

The segments selected by the first module are then passed to the second module, which performs the prediction or classification task. The modules are trained together, and the goal of training is to maximize both the score of the extracted segments and the accuracy of prediction or classification.

One of the data sets on which the researchers tested their system is a group of reviews from a website where users evaluate different beers. The data set includes the raw text of the reviews and the corresponding ratings, using a five-star system, on each of three attributes: aroma, palate, and appearance.

What makes the data attractive to natural-language-processing researchers is that it's also been annotated by hand, to indicate which sentences in the reviews correspond to which scores. For example, a review might consist of eight or nine sentences, and the annotator might have highlighted those that refer to the beer's "tan-colored head about half an inch thick," "signature Guinness smells," and "lack of carbonation." Each sentence is correlated with a different attribute rating.

Validation

As such, the data set provides an excellent test of the CSAIL researchers' system. If the first module has extracted those three phrases, and the second module has correlated them with the correct ratings, then the system has identified the same basis for judgment that the human annotator did.

In experiments, the system's agreement with the human annotations was 96 percent and 95 percent, respectively, for ratings of appearance and aroma, and 80 percent for the more nebulous concept of palate.

In the paper, the researchers also report testing their system on a database of free-form technical questions and answers, where the task is to determine whether a given question has been answered previously.

In unpublished work, they've applied it to thousands of pathology reports on breast biopsies, where it has learned to extract text explaining the bases for the pathologists' diagnoses. They're even using it to analyze mammograms, where the first module extracts sections of images rather than segments of text.

More information: Paper: "Rationalizing neural predictions" people.csail.mit.edu/taolei/pa … mnlp16_rationale.pdf

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Technique reveals the basis for machine-learning systems' decisions (2016, October 28) retrieved 26 April 2024 from https://techxplore.com/news/2016-10-technique-reveals-basis-machine-learning-decisions.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.