Technique sheds light on inner workings of neural nets trained to process language

September 11 2017, by Larry Hardesty

Researchers will present a new general-purpose technique for making sense of neural networks trained to perform natural-language-processing tasks, in which computers attempt to interpret freeform texts written in ordinary, or natural language (as opposed to a programming language, for example). Credit: Jose-Luis Olivares/MIT

Artificial-intelligence research has been transformed by machine-learning systems called neural networks, which learn how to perform tasks by analyzing huge volumes of training data.

During training, a neural net continually readjusts thousands of internal parameters until it can reliably perform some task, such as identifying objects in digital images or translating text from one language to another. But on their own, the final values of those parameters say very little about how the neural net does what it does.

Understanding what neural networks are doing can help researchers improve their performance and transfer their insights to other applications, and computer scientists have recently developed some clever techniques for divining the computations of particular neural networks.

But, at the 2017 Conference on Empirical Methods on Natural Language Processing starting this week, researchers from MIT's Computer Science and Artificial Intelligence Laboratory are presenting a new general-purpose technique for making sense of neural networks that are trained to perform natural-language-processing tasks, in which computers attempt to interpret freeform texts written in ordinary, or "natural," language (as opposed to a structured language, such as a database-query language).

The technique applies to any system that takes text as input and produces strings of symbols as output, such as an automatic translator. And because its analysis results from varying inputs and examining the effects on outputs, it can work with online natural-language-processing services, without access to the underlying software.

In fact, the technique works with any black-box text-processing system, regardless of its internal machinery. In their experiments, the researchers

show that the technique can identify idiosyncrasies in the work of human translators, too.

Theme and variations

The technique is analogous to one that has been used to analyze neural networks trained to perform computer vision tasks, such as object recognition. Software that systematically perturbs—or varies—different parts of an image and resubmits the image to an object recognizer can identify which image features lead to which classifications. But adapting that approach to natural language processing isn't straightforward.

"What does it even mean to perturb a sentence semantically?" asks Tommi Jaakkola, the Thomas Siebel Professor of Electrical Engineering and Computer Science at MIT and one of the new paper's two authors. "I can't just do a simple randomization. And what you are predicting is now a more complex object, like a sentence, so what does it mean to give an explanation?"

Somewhat ironically, to generate test sentences to feed to black-box neural nets, Jaakkola and David Alvarez-Melis, an MIT graduate student in electrical engineering and computer science and first author on the new paper, use a black-box neural net.

They begin by training a network to both compress and decompress natural sentences—to create some intermediate, compact digital representation of the sentence and then try to re-expand it into its original form. During training, the encoder and decoder are evaluated simultaneously, according to how faithfully the decoder's output matches the encoder's input.

Neural nets are intrinsically probabilistic: An object-recognition system fed an image of a small dog, for instance, might conclude that the image

has a 70 percent probability of representing a dog and a 25 percent probability of representing a cat. Similarly, Jaakkola and Alvarez-Melis' sentence-compressing network supplies alternatives for each word in a decoded sentence, along with the probabilities that each alternative is correct.

Because the network naturally uses the co-occurrence of words to increase its decoding accuracy, its output probabilities define a cluster of semantically related sentences. For instance, if the encoded sentence is "She gasped in surprise," the system might assign the alternatives "She squealed in surprise" or "She gasped in horror" as fairly high probabilities, but it would assign much lower probabilities to "She swam in surprise" or "She gasped in coffee."

For any sentence, then, the system can generate a list of closely related sentences, which Jaakkola and Alvarez-Melis feed to a black-box natural-language processor. The result is a long list of input-output pairs, which the researchers' algorithms can analyze to determine which changes to which inputs cause which changes to which outputs.

Test cases

The researchers applied their technique to three different set types of natural-language-processing system. One was a system that inferred words' pronunciation; another was a set of translators, two automated and one human; and the third was a simple computer dialogue system, which attempts to supply plausible responses to arbitrary remarks or questions.

As might be expected, the analysis of the translation systems demonstrated strong dependencies between individual words in the input and output sequences. One of the more intriguing results of that analysis, however, was the identification of gender biases in the texts on which

the machine translations systems were trained.

For instance, the nongendered English word "dancer" has two gendered translations in French, "danseur" and "danseuse." The system translated the sentence "The dancer is charming" using the feminine: "la danseuse est charmante." But the researchers' analysis showed that the choice of the word "danseuse" was as heavily influenced by the word "charming" as it was by the word "dancer." A different adjective might have resulted in a different translation of "dancer."

The dialogue system, which was trained on pairs of lines from Hollywood movies, was intentionally underpowered. Although the training set was large, the network itself was too small to take advantage of it.

"The other experiment we do is in flawed systems," Alvarez-Melis explains. "If you have a black-box model that is not doing a good job, can you first use this kind of approach to identify the problems? A motivating application of this kind of interpretability is to fix systems, to improve systems, by understanding what they're getting wrong and why."

In this case, the researchers' analyses showed that the dialogue system was frequently keying in on just a few words in an input phrase, which it was using to select a stock response—answering "I don't know" to any sentence that began with a query word such as "who" or "what," for example.

More information: A causal framework for explaining the predictions of black-box sequence-to-sequence models. people.csail.mit.edu/tommi/pap … AlvJaa_EMNLP2017.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