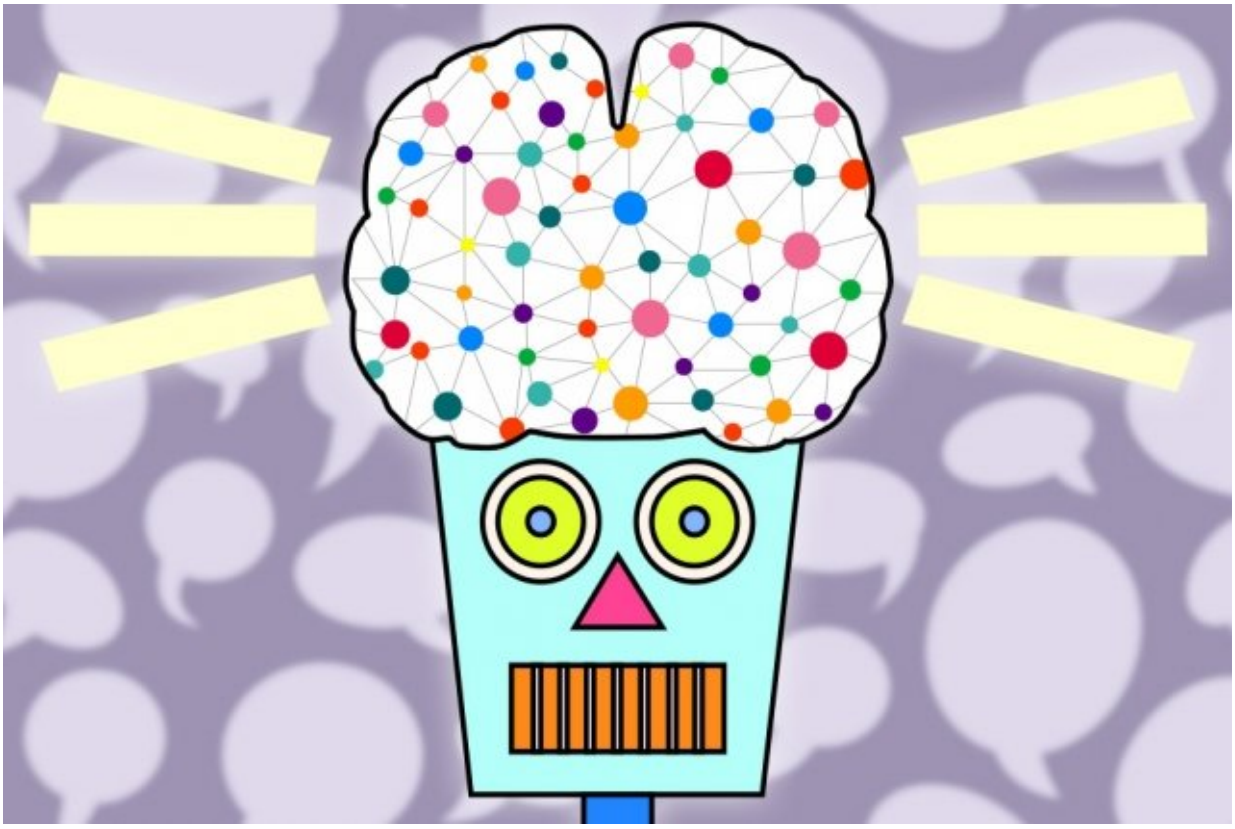
Technique illuminates the inner workings of artificial-intelligence systems that process language

December 11 2017, by Larry Hardesty



Neural nets are so named because they roughly approximate the structure of the human brain. Typically, they're arranged into layers, and each layer consists of many simple processing units — nodes — each of which is connected to several nodes in the layers above and below. Data is fed into the lowest layer, whose nodes process it and pass it to the next layer. The connections between layers have different "weights," which determine how much the output of any one node

figures into the calculation performed by the next. Credit: Chelsea Turner/MIT

Neural networks, which learn to perform computational tasks by analyzing huge sets of training data, have been responsible for the most impressive recent advances in artificial intelligence, including speech-recognition and automatic-translation systems.

During training, however, a neural net continually adjusts its internal settings in ways that even its creators can't interpret. Much recent work in computer science has focused on clever techniques for determining just how neural nets do what they do.

In several recent papers, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and the Qatar Computing Research Institute have used a recently developed interpretive technique, which had been applied in other areas, to analyze neural networks trained to do machine translation and speech recognition.

They find empirical support for some common intuitions about how the networks probably work. For example, the systems seem to concentrate on lower-level tasks, such as sound recognition or part-of-speech recognition, before moving on to higher-level tasks, such as transcription or semantic interpretation.

But the researchers also find a surprising omission in the type of data the translation network considers, and they show that correcting that omission improves the network's performance. The improvement is modest, but it points toward the possibility that analysis of neural networks could help improve the accuracy of artificial intelligence systems.

"In machine translation, historically, there was sort of a pyramid with different layers," says Jim Glass, a CSAIL senior research scientist who worked on the project with Yonatan Belinkov, an MIT graduate student in electrical engineering and computer science. "At the lowest level there was the word, the surface forms, and the top of the pyramid was some kind of interlingual representation, and you'd have different layers where you were doing syntax, semantics. This was a very abstract notion, but the idea was the higher up you went in the pyramid, the easier it would be to translate to a new language, and then you'd go down again. So part of what Yonatan is doing is trying to figure out what aspects of this notion are being encoded in the network."

The work on machine translation was presented recently in two papers at the International Joint Conference on Natural Language Processing. On one, Belinkov is first author, and Glass is senior author, and on the other, Belinkov is a co-author. On both, they're joined by researchers from the Qatar Computing Research Institute (QCRI), including Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and Stephan Vogel. Belinkov and Glass are sole authors on the paper analyzing speech recognition systems, which Belinkov presented at the Neural Information Processing Symposium last week.

Leveling down

Neural nets are so named because they roughly approximate the structure of the human brain. Typically, they're arranged into layers, and each layer consists of many simple processing units—nodes—each of which is connected to several nodes in the layers above and below. Data are fed into the lowest layer, whose nodes process it and pass it to the next layer. The connections between layers have different "weights," which determine how much the output of any one node figures into the calculation performed by the next.

During training, the weights between nodes are constantly readjusted. After the network is trained, its creators can determine the weights of all the connections, but with thousands or even millions of nodes, and even more connections between them, deducing what algorithm those weights encode is nigh impossible.

The MIT and QCRI researchers' technique consists of taking a trained network and using the output of each of its layers, in response to individual training examples, to train another neural network to perform a particular task. This enables them to determine what task each layer is optimized for.

In the case of the speech recognition network, Belinkov and Glass used individual layers' outputs to train a system to identify "phones," distinct phonetic units particular to a spoken language. The "t" sounds in the words "tea," "tree," and "but," for instance, might be classified as separate phones, but a speech recognition system has to transcribe all of them using the letter "t." And indeed, Belinkov and Glass found that lower levels of the network were better at recognizing phones than higher levels, where, presumably, the distinction is less important.

Similarly, in an earlier paper, presented last summer at the Annual Meeting of the Association for Computational Linguistics, Glass, Belinkov, and their QCRI colleagues showed that the lower levels of a machine-translation network were particularly good at recognizing parts of speech and morphology—features such as tense, number, and conjugation.

Making meaning

But in the new paper, they show that higher levels of the network are better at something called semantic tagging. As Belinkov explains, a part-of-speech tagger will recognize that "herself" is a pronoun, but the

meaning of that pronoun—its semantic sense—is very different in the sentences "she bought the book herself" and "she herself bought the book." A semantic tagger would assign different tags to those two instances of "herself," just as a machine translation system might find different translations for them in a given target language.

The best-performing machine-translation networks use so-called encoding-decoding models, so the MIT and QCRI researchers' network uses it as well. In such systems, the input, in the source language, passes through several layers of the network—known as the encoder—to produce a vector, a string of numbers that somehow represent the semantic content of the input. That vector passes through several more layers of the network—the decoder—to yield a translation in the target language.

Although the encoder and decoder are trained together, they can be thought of as separate networks. The researchers discovered that, curiously, the lower layers of the encoder are good at distinguishing morphology, but the higher layers of the decoder are not. So Belinkov and the QCRI researchers retrained the network, scoring its performance according to not only accuracy of translation but also analysis of morphology in the target language. In essence, they forced the decoder to get better at distinguishing morphology.

Using this technique, they retrained the network to translate English into German and found that its accuracy increased by 3 percent. That's not an overwhelming improvement, but it's an indication that looking under the hood of neural networks could be more than an academic exercise.

More information: Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems. arxiv.org/abs/1709.04482

Evaluating Layers of Representation in Neural Machine Translation on

Part-of-Speech and Semantic Tagging Tasks. people.csail.mit.edu/belinkov/ … lp2017-semantics.pdf

Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder: people.csail.mit.edu/belinkov/ … cnlp2017-decoder.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 illuminates the inner workings of artificial-intelligence systems that process language (2017, December 11) retrieved 27 April 2024 from https://techxplore.com/news/2017-12-technique-illuminates-artificial-intelligence-language.html