Putting neural networks under the microscope

February 1 2019, by Rob Matheson

Researchers from MIT and the Qatar Computing Research Institute (QCRI) are putting the machine-learning systems known as neural networks under the microscope. Credit: MIT News

Researchers from MIT and the Qatar Computing Research Institute

(QCRI) are putting the machine-learning systems known as neural networks under the microscope.

In a study that sheds light on how these systems manage to translate text from one language to another, the researchers developed a method that pinpoints individual nodes, or "neurons," in the networks that capture specific linguistic features.

Neural networks learn to perform computational tasks by processing huge sets of training data. In machine translation, a network crunches language data annotated by humans, and presumably "learns" linguistic features, such as word morphology, sentence structure, and word meaning. Given new text, these networks match these learned features from one language to another, and produce a translation.

But, in training, these networks basically adjust internal settings and values in ways the creators can't interpret. For machine translation, that means the creators don't necessarily know which linguistic features the network captures.

In a paper being presented at this week's Association for the Advancement of Artificial Intelligence conference, the researchers describe a method that identifies which neurons are most active when classifying specific linguistic features. They also designed a toolkit for users to analyze and manipulate how their networks translate text for various purposes, such as making up for any classification biases in the training data.

In their paper, the researchers pinpoint neurons that are used to classify, for instance, gendered words, past and present tenses, numbers at the beginning or middle of sentences, and plural and singular words. They also show how some of these tasks require many neurons, while others require only one or two.

"Our research aims to look inside neural networks for language and see what information they learn," says co-author Yonatan Belinkov, a postdoc in the Computer Science and Artificial Intelligence Laboratory (CSAIL). "This work is about gaining a more fine-grained understanding of neural networks and having better control of how these models behave."

Co-authors on the paper are: senior research scientist James Glass and undergraduate student Anthony Bau, of CSAIL; and Hassan Sajjad, Nadir Durrani, and Fahim Dalvi, of QCRI.

Putting a microscope on neurons

Neural networks are structured in layers, where each layer consists of many processing nodes, each connected to nodes in layers above and below. Data are first processed in the lowest layer, which passes an output to the above layer, and so on. Each output has a different "weight" to determine how much it figures into the next layer's computation. During training, these weights are constantly readjusted.

Neural networks used for machine translation train on annotated language data. In training, each layer learns different "word embeddings" for one word. Word embeddings are essentially tables of several hundred numbers combined in a way that corresponds to one word and that word's function in a sentence. Each number in the embedding is calculated by a single neuron.

In their past work, the researchers trained a model to analyze the weighted outputs of each layer to determine how the layers classified any given embedding. They found that lower layers classified relatively simpler linguistic features—such as the structure of a particular word—and higher levels helped classify more complex features, such as how the words combine to form meaning.

In their new work, the researchers use this approach to determine how learned word embeddings make a linguistic classification. But they also implemented a new technique, called "linguistic correlation analysis," that trains a model to home in on the individual neurons in each word embedding that were most important in the classification.

The new technique combines all the embeddings captured from different layers—which each contain information about the word's final classification—into a single embedding. As the network classifies a given word, the model learns weights for every neuron that was activated during each classification process. This provides a weight to each neuron in each word embedding that fired for a specific part of the classification.

"The idea is, if this neuron is important, there should be a high weight that's learned," Belinkov says. "The neurons with high weights are the ones more important to predicting the certain linguistic property. You can think of the neurons as a lot of knobs you need to turn to get the correct combination of numbers in the embedding. Some knobs are more important than others, so the technique is a way to assign importance to those knobs."

Neuron ablation, model manipulation

Because each neuron is weighted, it can be ranked in order of importance. To that end, the researchers designed a toolkit, called NeuroX, that automatically ranks all neurons of a neural network according to their importance and visualizes them in a web interface.

Users upload a network they've already trained, as well as new text. The app displays the text and, next to it, a list of specific neurons, each with an identification number. When a user clicks on a neuron, the text will be highlighted depending on which words and phrases the neuron

activates for. From there, users can completely knock out—or "ablate"—the neurons, or modify the extent of their activation, to control how the network translates.

The task of ablation was used to determine if the researchers' method accurately pinpointed the correct high-ranking neurons. In their paper, the researchers used the tool to show that, by ablating high ranking neurons in a network, its performance in classifying correlated linguistic features dipped significantly. Alternatively, when they ablated lower-ranking neurons, performance suffered, but not as dramatically.

"After you get all these rankings, you want to see what happens when you kill these neurons and see how badly it affects performance," Belinkov says. "That's an important result proving that the neurons we find are, in fact, important to the classification process."

One interesting application for the toolkit is helping limit biases in language data. Machine-translation models, such as Google Translate, may train on data with gender bias, which can be problematic for languages with gendered words. Certain professions, for instance, may be more often referred to as male, and others as female. When a network translates new text, it may only produce the learned gender for those words. In many online English-to-Spanish translations, for instance, "doctor" often translates into its masculine version, while "nurse" translates into its feminine version.

"But we find we can trace individual neurons in charge of linguistic properties like gender," Belinkov says. "If you're able to trace them, maybe you can intervene somehow and influence the translation to translate these words more to the opposite gender … to remove or mitigate the bias."

In preliminary experiments, the researchers modified neurons in a

network to change translated text from past to present tense with 67 percent accuracy. They ablated to switch the gender of the words with 21 percent accuracy. "It's still a work in progress," Belinkov says. A next step, he adds, is fine-tuning the web application to achieve more accurate ablation and manipulation.

More information: What Is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models: arXiv:1812.09355 [cs.CL] arxiv.org/abs/1812.09355

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