Popular Science Abstract

Imagine that it is the last day before a very important exam you have been diligently preparing for a long time. You read the last page in your book, close it and go to sleep. Now imagine your surprise when during the exam, you realize that the only thing you can remember is the content of the last page you read the day before while rehearsing for the exam.

It turns out that your knowledge has been catastrophically forgotten -- a phenomenon that has plagued artificial neural networks for several decades. Catastrophic forgetting prevents neural networks from gradually acquiring knowledge because it causes all the previously acquired knowledge contained in the neural network to be replaced by the information the neural network is learning at that moment.

How scientists teaching neural networks have dealt with this problem leaves much desired and poses many difficulties. To effectively teach the neural network the content of one book, we would have to tear out all the book's pages and then randomly, without regard to page numbering, teach the network the information contained on each page. If we repeat this activity many times, ensuring that each page is read by the network a few dozen times, we can be sure that the neural network has learned the material contained in a given book.

There are three major drawbacks to this approach. First, we have just destroyed the book. Second, we need access to the entire book at any point in the neural network learning process, and books can be heavy. Third, because the pages of the book were torn out and randomly mixed, it may happen that the neural network will try to understand the material from the very end of the book, not yet understanding the basics presented at the very beginning.

We can imagine how inefficient the current process of learning neural networks is. Unfortunately, despite many attempts, scientists have failed to identify the cause, let alone solve the phenomenon of catastrophic forgetting. In this project, we will analyze the effect of the task that a neural network solves on the forgetting phenomenon. As our previous work shows, forgetting is more severe when the neural network learns, for example, to discriminate between individual objects. In contrast, surprisingly, the forgetting phenomenon is much weaker when the network learns to redraw the same objects. Our study suggests that a network that learns to redraw objects creates an internal representation of the redrawn object that is significantly more durable and resistant to forgetting than the representation of objects created when the neural network is tasked with discriminating them.

In further work, we plan to analyze the effect of different types of tasks on the internal representations created by the neural network and their durability and resistance to the phenomenon of forgetting. The results of our research will allow us to better understand the phenomenon of forgetting and create methods to reduce this phenomenon.

We hope that we will be able to lend our books to neural networks without any fear and get back an intact copy soon.