The New York Times | https://www.nytimes.com/2021/07/20/technology/ai-education-neural-networks.html

Can A.I. Grade Your Next Test?

Neural networks could give online education a boost by providing automated feedback to students.

By Cade Metz

July 20, 2021

This spring, Philips Pham was among the more than 12,000 people in 148 countries who took an online class called Code in Place. Run by Stanford University, the course taught the fundamentals of computer programming.

Four weeks in, Mr. Pham, a 23-year-old student living at the southern tip of Sweden, typed his way through the first test, trying to write a program that could draw waves of tiny blue diamonds across a black-and-white grid. Several days later, he received a detailed critique of his code.

It applauded his work, but also pinpointed an error. “Seems like you have a small mistake,” the critique noted. “Perhaps you are running into the wall after drawing the third wave.”

The feedback was just what Mr. Pham needed. And it came from a machine.

During this online class, a new kind of artificial intelligence offered feedback to Mr. Pham and thousands of other students who took the same test. Built by a team of Stanford researchers, this automated system points to a new future for online education, which can so easily reach thousands of people but does not always provide the guidance that many students need and crave.

“We’ve deployed this in the real world, and it works better than we expected,” said Chelsea Finn, a Stanford professor and A.I. researcher who helped build the new system.

Dr. Finn and her team designed this system solely for Stanford’s programming class. But they used techniques that could automate student feedback in other situations, including for classes beyond programming.

Oren Etzioni, chief executive of the Allen Institute for Artificial Intelligence and a former professor of computer science at the University of Washington, cautioned that these techniques are a very long way from duplicating human instructors. Feedback and advice from professors, teaching assistants and tutors is always preferable to an automated critique.

Still, Dr. Etzioni called the Stanford project a "step in an important direction," with automated feedback better than none at all.

The online course taken by Mr. Pham and thousands of others this spring is based on a class that Stanford has offered for more than a decade. Each semester, the university gives students a midterm test filled with programming exercises, and it keeps a digital record of the results, including the reams of code written by students as well as pointed critiques of each program from university instructors.

This decade of data is what drove the university's new experiment in artificial intelligence.

Dr. Finn and her team built a neural network, a mathematical system that can learn skills from vast amounts of data. By pinpointing patterns in thousands of cat photos, a neural network can learn to identify a cat. By analyzing hundreds of old phone calls, it can learn to recognize spoken words. Or, by examining the way teaching assistants evaluate coding tests, it can learn to evaluate these tests on its own.

The Stanford system spent hours analyzing examples from old midterms, learning from a decade of possibilities. Then it was ready to learn more. When given just a handful of extra examples from the new exam offered this spring, it could quickly grasp the task at hand.

"It sees many kinds of problems," said Mike Wu, another researcher who worked on the project. "Then it can adapt to problems it has never seen before."

This spring, the system provided 16,000 pieces of feedback, and students agreed with the feedback 97.9 percent of the time, according to a study by the Stanford researchers. By comparison, students agreed with the feedback from human instructors 96.7 percent of the time.

Mr. Pham, an engineering student at Lund University in Sweden, was surprised the technology worked so well. Although the automated tool was unable to evaluate one of his programs (presumably because he had written a snippet of code unlike anything the A.I. had ever seen), it both identified specific bugs in his code, including what is known in computer programming and mathematics as a fence post error, and suggested ways of fixing them. "It is seldom you receive such well thought out feedback," Mr. Pham said.

The technology was effective because its role was so sharply defined. In taking the test, Mr. Pham wrote code with very specific aims, and there were only so many ways that he and other students could go wrong.

But given the right data, neural networks can learn a range of tasks. This is the same fundamental technology that identifies faces in the photos you post to Facebook, recognizes the commands you bark into your iPhone and translates from one language to another on services like Skype and Google Translate. For the Stanford team and other researchers, the hope is that these techniques can automate education in many other ways.

Researchers have been building automated teaching tools since the 1970s, including robo-tutors and computerized essay graders. But progress has been slow. Building a system that can simply and clearly guide students often requires years of work, with designers struggling to define each tiny piece of behavior.

Using the methods that drove the Stanford project, researchers can significantly accelerate this work. “There is real power in data,” said Peter Foltz, a professor at the University of Colorado who has spent decades developing systems that can automatically grade prose essays. “As machines get more examples, they can generalize.”

Prose may seem very different from computer code. But in this case, it is not. In recent years, researchers have built technology that can analyze natural language in much the same way the Stanford system analyzes computer code.

Although the Stanford system provides sharp feedback, it is useless if students have any questions about where they went wrong. But for Chris Piech, the Stanford professor who helped oversee the class, replacing instructors is not the goal.

The new automated system is a way of reaching more students than instructors could otherwise reach on their own. And if it can clearly pinpoint problems in student code, showing the specific coding mistakes they are making and how frequently they are making them, it could help instructors better understand which students need help and how to help them. As Dr. Piech put it: “The future is symbiotic — teachers and A.I. working together.”