Why Google believes machine learning is its future

Enlarge / Google CEO Sundar Pichai speaks during the Google I/O Developers Conference on May 7, 2019.
David Paul Morris/Bloomberg via Getty Images

reader comments

83 with 60 posters participating

Share this story

One of the most interesting demos at this week's Google I/O keynote featured a new version of Google's voice assistant that's due out later this year. A Google employee asked the Google Assistant to bring up her photos and then show her photos with animals. She tapped one and said, "Send it to Justin." The photo was dropped into the messaging app.

From there, things got more impressive.

"Hey Google, send an email to Jessica," she said. "Hi Jessica, I just got back from Yellowstone and completely fell in love with it." The phone transcribed her words, putting "Hi Jessica" on its own line.

"Set subject to Yellowstone adventures," she said. The assistant understood that it should put "Yellowstone adventures" into the subject line, not the body of the message.

Then without any explicit command, the woman went back to dictating the body of the message. Finally she said "send it," and Google's assistant did.

[embedded content]
Google is also working to expand the assistant's understanding of personal references, the company said. If a user says, "Hey Google, what's the weather like at Mom's house," Google will be able to figure out that "mom's house" refers to the home of the user's mother, look up her address, and provide a weather forecast for her city.

Google says that its next-generation assistant is coming to "new Pixel phones"—that is, the phones that come after the current Pixel 3 line—later this year.

Obviously, there's a big difference between a canned demo and a shipping product. We'll have to wait and see if typical interactions with the new assistant work this well. But Google seems to be making steady progress toward the dream of building a virtual assistant that can competently handle even complex tasks by voice.

A lot of the announcements at I/O were like this: not the announcement of major new products, but the use of machine learning techniques to gradually make a range of Google products more sophisticated and helpful. Google also touted a number of under-the-hood improvements to its machine learning software, which will allow both Google-created and third-party software to use more sophisticated machine learning techniques.

In particular, Google is making a big push to shift machine learning operations from the cloud onto peoples' mobile devices. This should allow ML-powered applications to be faster, more private, and able to operate offline.

Google has led the charge on machine learning

Enlarge / A circuit board containing Google’s tensor processor unit.

If you ask machine learning experts when the current deep learning boom started, many will point to a 2012 paper known as “AlexNet” after lead author Alex Krizhevsky. The authors, a trio of researchers from the University of Toronto, entered the ImageNet competition to classify images into one of a thousand categories.

The ImageNet organizers supplied more than a million labeled example images to train the networks. AlexNet achieved unprecedented accuracy by using a deep neural network, with eight trainable layers and 650,000 neurons. They were able to train such a massive network on so much data because they figured out how to harness consumer-grade GPUs, which are designed for large-scale parallel processing.

AlexNet demonstrated the importance of what you might call the three-legged stool of deep learning: better algorithms, more training data, and more computing power. Over the last seven years, companies have been scrambling to beef up their capabilities on all three fronts, resulting in better and better performance.

Google has been leading this charge almost from the beginning.

Two years after AlexNet won an image recognition competition called ImageNet in 2012, Google entered the contest with an even deeper neural network and took top prize. The company has hired dozens of top-tier machine learning experts, including the 2014 acquisition of deep learning startup DeepMind, keeping the company at the forefront of neural network design.

The company also has unrivaled access to large data sets. A 2013 paper described how Google was using deep neural networks to recognize address numbers in tens of millions of images captured by Google Street View.

Google has been hard at work on the hardware front, too. In 2016, Google announced that it had created a custom chip called a Tensor Processing Unit specifically designed to accelerate the operations used by neural networks.

“Although Google considered building an Application-Specific Integrated Circuit (ASIC) for neural networks as early as 2006, the situation became urgent in 2013,” Google wrote in 2017. “That’s when we realized that the fast-growing computational demands of neural networks could require us to double the number of data centers we operate.”

This is why Google I/O has had such a focus on machine learning for the last three years. The company believes that these assets—a small army of machine learning experts, vast amounts of data, and its own custom silicon—make it ideally positioned to exploit the opportunities presented by machine learning.

This year’s Google I/O didn’t actually have a lot of major new ML-related product announcements because the company has already baked machine learning into many of its major products. Android has had voice recognition and the Google Assistant for years. Google Photos has long had an impressive ML-based search function. Last year, Google introduced Google Duplex, which makes a reservation on behalf of a user with an

uncannily realistic human voice created by software.

Instead, I/O presentations on machine learning focused on two areas: shifting more machine learning activity onto smartphones and using machine learning to help disadvantaged people—including people who are deaf, illiterate, or suffering from cancer.

Squeezing machine learning onto smartphones

Justin Sullivan/Getty Images
Past efforts to make neural networks more accurate have involved making them deeper and more complicated. This approach has produced impressive results, but it has a big downside: the networks often wind up being too complex to run on smartphones.

People have mostly dealt with this by offloading computation to the cloud. Early versions of Google and Apple's voice assistants would record audio and upload it to the companies

servers for processing. That worked all right, but it had three significant downsides: it had higher latency, it had weaker privacy protection, and the feature would only work online.

So Google has been working to shift more and more computation on-device. Current Android devices already have basic on-device voice recognition capabilities, but Google's virtual assistant requires an Internet connection. Google says that situation will change later this year with a new offline mode for Google Assistant.

This new capability is a big reason for the lightning-fast response times demonstrated by this week's demo. Google says the assistant will be "up to 10 times faster" for certain tasks.

https://arstechnica.com/?p=1502677