New open-source platform cuts costs for running AI

December 7 2023, by Patricia Waldron

Credit: CC0 Public Domain

Cornell researchers have released a new, open-source platform called Cascade that can run artificial intelligence (AI) models in a way that slashes expenses and energy costs while dramatically improving

performance.

Cascade is designed for settings like smart traffic intersections, medical diagnostics, equipment servicing using augmented reality, digital agriculture, smart power grids and automatic product inspection during manufacturing—situations where AI models must react within a fraction of a second. It is already in use by College of Veterinary Medicine researchers monitoring cows for risk of mastitis.

With the rise of AI, many companies are eager to leverage new capabilities but worried about the associated computing costs and the risks of sharing private data with AI companies or sending sensitive information into the cloud—far-off servers accessed through the internet.

Also, today's AI models are slow, limiting their use in settings where data must be transferred back and forth or the model is controlling an automated system. A team led by Ken Birman, professor of computer science in the Cornell Ann S. Bowers College of Computing and Information Science, combined several innovations to address these concerns.

Birman partnered with Weijia Song, a senior research associate, to develop an edge computing system they named Cascade. Edge computing is an approach that places the computation and data storage closer to the sources of data, protecting sensitive information. Song's "zero copy" edge computing design minimizes data movement. The AI models don't have to wait to fetch data when reacting to an event, which enables faster responses, the researchers said.

"Cascade enables users to put machine learning and data fusion really close to the edge of the internet, so artificially intelligent actions can occur instantly," Birman said. "This contrasts with standard cloud

computing approaches, where the frequent movement of data from machine to machine forces those same AIs to wait, resulting in long delays perceptible to the user."

Cascade is giving impressive results, with most programs running two to 10 times faster than cloud-based applications, and some computer vision tasks speeding up by factors of 20 or more. Larger AI models see the most benefit.

Moreover, the approach is easy to use: "Cascade often requires no changes at all to the AI software," Birman said.

Alicia Yang, a doctoral student in the field of computer science, was one of several student researchers in the effort. She developed Navigator, a memory manager and task scheduler for AI workflows that further boosts performance. "Navigator really pays off when a number of applications need to share expensive hardware," Yang said. "Compared to cloud-based approaches, Navigator accomplishes the same work in less time and uses the hardware far more efficiently."

In CVM, Parminder Basran, associate research professor of medical oncology in the Department of Clinical Sciences, and Matthias Wieland, assistant professor in the Department of Population Medicine and Diagnostic Sciences, are using Cascade to monitor dairy cows for signs of increased mastitis—a common infection in the mammary gland that reduces milk production.

By imaging the udders of thousands of cows during each milking session and comparing the new photos to those from past milkings, an AI model running on Cascade identifies dry skin, open lesions, rough teat ends and other changes that may signal disease. If early symptoms are detected, cows could be subjected to a medicinal rinse at the milking station to potentially head off a full-blown infection.

Thiago Garrett, a visiting researcher from the University of Oslo, used Cascade to build a prototype "smart traffic intersection." His solution tracks crowded settings packed with people, cars, bicycles and other objects, anticipates possible collisions and warns of risks—within milliseconds after images are captured. When he ran the same AI model on a cloud computing infrastructure, it took seconds to sense possible accidents, far too late to sound a warning.

With the new open-source release, Birman's group hopes other researchers will explore possible uses for Cascade, making AI applications more widely accessible.

"Our goal is to see it used," Birman said. "This open-source release will allow the public to benefit from what we created."

Provided by Cornell University

Citation: New open-source platform cuts costs for running AI (2023, December 7) retrieved 28 April 2024 from https://techxplore.com/news/2023-12-open-source-platform-ai.html