Using quantum computing to protect AI from attack

August 14 2023, by Max West and Sarah Erfani

Credit: AI-generated image (disclaimer)

Machine learning is a field of artificial intelligence (AI) where computer models become experts in various tasks by consuming large amounts of data. This is instead of a human explicitly programming this level of expertise.

For example, modern chess AIs do not need to be taught chess strategies by human grandmasters, but can "learn" them independently by playing millions of games against copies of themselves.

This is invaluable in situations where writing down explicit instructions is impractical, if not impossible—how do you define a mathematical function that can tell you if a picture contains a cat or a dog?

Human children never learn any such function, but rather see many examples of cats and dogs, then eventually develop an understanding of their differences.

Machine learning is about replicating this process in computers.

But despite their incredible successes and increasingly widespread deployment, machine learning-based frameworks remain highly susceptible to adversarial attacks—that is, malicious tampering with their data causing them to fail in surprising ways.

For example, image-classifying models (which analyze photos to identify and recognize a wide variety of criteria) can often be fooled by the addition of well-crafted alterations (known as perturbations) to their input images that are so small they are imperceptible to the human eye. And this can be exploited.

The continued vulnerability to attacks like these also raises serious questions about the safety of deploying machine learning neural networks in potentially life-threatening situations. This includes applications like self-driving cars, where the system could be confused into driving through an intersection by an innocuous piece of graffiti on a stop sign.

At a crucial time when the development and deployment of AI are

rapidly evolving, our research team is looking at ways we can use quantum computing to protect AI from these vulnerabilities,

Machine learning and quantum computing

Recent advances in quantum computing have generated much excitement about the prospect of enhancing machine learning with quantum computers. Various "quantum machine learning" algorithms already having been proposed, including quantum generalizations of the standard classical methods.

Generalization refers to a learning model's ability to adapt properly to new, previously unseen data.

It is believed quantum machine learning models can learn certain types of data drastically faster than any model designed for current or "classical" computers.

Ordinary computers work with bits of data that can be either "zero" or "one"—a two-level classical system.

Quantum computers work with "qubits," states of two-level quantum systems, which exhibit strange additional properties that can be harnessed in order to tackle certain problems more efficiently than their classical counterparts

What is less clear, however, is how widespread these speedups will be and how useful quantum machine learning will be in practice.

This is because although quantum computers are expected to efficiently learn a wider class of models than their classical counterparts, there's no guarantee these new models will be useful for most machine-learning tasks in which people are actually interested. These might include

medical classification problems or generative AI systems.

These challenges motivated our team to consider what other benefits quantum computing could bring to machine learning tasks—other than the usual goals of improving efficiency or accuracy.

Shielding AI from attacks

In our latest work, published in *Physical Review Research*, we suggest quantum machine learning models may be better defended against adversarial attacks generated by classical computers.

Adversarial attacks work by identifying and exploiting the features used by a machine learning model.

But the features used by generic quantum machine learning models are inaccessible to classical computers, and therefore invisible to an adversary armed only with classical computing resources.

These ideas could also be used to detect the presence of adversarial attacks, by simultaneously using classical and quantum networks.

Under normal conditions, both networks should make the same predictions, but in the presence of an attack—their outputs will diverge.

While this is encouraging, quantum machine learning continues to face significant challenges. Chief among them is the massive capability gap that separates classical and quantum computing hardware.

Today's quantum computers remain significantly limited by their size and their high error rates, which preclude them from carrying out long calculations.

Formidable engineering challenges remain, but if these can be overcome, the unique capabilities of large-scale quantum computers will doubtless deliver surprising benefits across a wide range of fields.

More information: Maxwell T. West et al, Benchmarking adversarially robust quantum machine learning at scale, *Physical Review Research* (2023). DOI: 10.1103/PhysRevResearch.5.023186

Provided by University of Melbourne