Taking a quantum leap

22/04/2022 More and more vehicle functions are based on artificial intelligence. However, conventional processors and even graphics chips are increasingly reaching their limits when it comes to calculations required for neural networks. Porsche Engineering reports on new technologies that will speed up AI calculations in the future.

Artificial intelligence (AI) is a key technology for the automotive industry—and fast hardware is correspondingly important for the complex back-end calculations involved. After all, it will only be possible to bring new functions into series production in the future with high-performance computers. “Autonomous driving is one of the most demanding AI applications of all,” explains Dr. Joachim Schaper, Senior Manager AI and Big Data at Porsche Engineering. “The algorithms learn from a multitude of examples collected by test vehicles using cameras, radar, or other sensors in real traffic.”

Conventional data centers are increasingly unable to cope with the growing demands. “It now takes days to train a single variant of a neural network,” explains Schaper. So in his view, one thing is clear: Car manufacturers need new technologies for AI calculations that can help the algorithms learn much faster. To achieve this, as many vector-matrix multiplications as possible must be executed in parallel in the complex deep neural networks (DNNs)—a task in which graphics processing units (GPUs)

specialize. Without them, the amazing advances in AI in recent years would not have been possible.

50 times the size of a GPU

Graphics cards were not originally designed for AI use, however, but to process image data as efficiently as possible. They are increasingly stretched to the limit when it comes to training algorithms for autonomous driving. Hardware specialized in AI is therefore required for even faster calculations. The Californian company Cerebras has presented a possible solution. Their Wafer Scale Engine (WSE) is optimally tailored to the requirements of neural networks by combining as much computing power as possible on one giant computer chip. It is more than 50 times the size of a normal graphics processor and offers space for 850,000 computing cores—over 100 times as many as on a current top GPU.

In addition, Cerebras engineers have networked the computational cores together with high-bandwidth data lines. According to the manufacturer, the network on the Wafer Scale Engine transports 220 petabits per second. Cerebras has also widened the bottleneck within the GPUs: Data travels between memory and computing unit nearly 10,000 times faster than in high-performance GPUs—at 20 petabytes per second.

To save even more time, Cerebras mimics a trick of the brain. There, neurons work only when they get signals from other neurons. The many connections that are currently inactive do not need any resources. In DNNs, on the other hand, vector-matrix multiplication often involves multiplying by the number zero. This costs time unnecessarily. The Wafer Scale Engine therefore refrains from doing so. “All zeros are filtered out,” Cerebras writes in its white paper on the WSE. So the chip only performs operations that produce a non-zero result.

One drawback of the chip is its high electrical power requirement of 23 kW and requires water cooling. Cerebras has therefore developed its own server housing for use in data centers. The Wafer Scale Engine is already being tested in the data centers of some research institutes. AI expert Joachim Schaper believes the giant chip from California could also accelerate automotive development. “By using this chip, a week’s training could theoretically be reduced to just a few hours,” he estimates. “However, the technology has yet to prove that in practical tests.”

Light instead of electrons

As unusual as the new chip is, like its conventional predecessors it also works with conventional transistors. Companies like Boston-based Lightelligence and Lightmatter want to use the much faster medium of light for AI calculations instead of comparatively slow electronics, and are building optical chips to do so. DNNs could thus work “at least several hundred times faster than electronic ones,” write developers at Lightelligence.

To do this, Lightelligence and Lightmatter use the phenomenon of interference. When light waves

amplify or cancel each other, they form a light-dark pattern. If you direct the interference in a certain way, the new pattern corresponds to the vector-matrix multiplication of the old pattern. So the light waves can "do math." To make this practical, the Boston developers etched tiny light guides into a silicon chip. Like in a textile fabric, they cross each other several times. Interference takes place at the crossings. In between, tiny heating elements regulate the refractive index of the light guide, allowing the light waves to be shifted against each other. This makes it possible to control their interference and perform vector-matrix multiplications.

However, the Boston companies do not dispense with electronics altogether. They combine their light computers with conventional electronic components that store data and perform all calculations except vector-matrix multiplications. These include, for example, the nonlinear activation functions that modify the output values of each neuron before they move on to the next layer.

With the combination of optical and digital computing, DNNs can be computed extremely quickly. "Their main advantage is low latency," explains Lindsey Hunt, a spokesperson for Lightelligence. For example, this allows the DNN to detect objects in images faster, such as pedestrians and e-scooter riders. In autonomous driving, this could lead to faster reactions in critical situations. "In addition, the optical system makes more decisions per watt of electrical energy," Hunt said. That's especially important as increasing computing power in vehicles increasingly comes at the expense of fuel economy and range.

The solutions from Lightmatter and Lightelligence can be inserted as modules into conventional computers to speed up AI computations—much like graphics cards. In principle, they could also be integrated into vehicles, for example to implement autonomous driving functions. "Our technology is very well suited to serve as an inference engine for an autonomous car," explains Lindsey Hunt. AI expert Schaper has a similar view: "If Lightelligence succeeds in building components suitable for automobiles, this could greatly accelerate the introduction of complex AI functions in vehicles." The technology is now ready for the market: The company is planning its first pilot tests with customers in the year 2022.

The quantum computer as an AI turbo

Quantum computers are somewhat further away from practical application. They, too, will accelerate AI calculations because they can process vast amounts of data in parallel. To do this, they work with so-called "qubits." Unlike the classical unit of information, the bit, a qubit can represent the two binary values 0 and 1 simultaneously. The two numbers coexist in a superposition state that is only possible in quantum mechanics.

Quantum computers could turbocharge artificial intelligence when if comes to classifying things, for example in traffic. There are many different categories of objects there, including bicycles, cars, pedestrians, signs, wet and dry roads. They differ in terms of many properties, which is why experts talk about "pattern recognition in higher-dimensional spaces."

“The more complicated the patterns, the harder it is for conventional computers to distinguish classes,” explains Heike Riel, who heads IBM’s quantum research in Europe and Africa. That’s because with each dimension, it becomes more costly to calculate the similarity of two objects: How similar are an e-scooter rider and a rollator user trying to cross the street? Quantum computers can work efficiently in high-dimensional spaces compared to conventional computers. For certain problems, this property could be useful and result in some problems being solved faster with the help of quantum computers than with conventional high-performance computers.

IBM researchers have analyzed statistical models that can be trained for data classification. Initial results suggest that cleverly chosen quantum models work better than conventional methods for certain datasets. The quantum models are easier to train and appear to have greater capacity—allowing them to learn more complicated relationships.

Riel admits that while today’s quantum computers can be used to test these algorithms, they do not yet have an advantage over conventional computers. However, the development of quantum computers is progressing rapidly. Both the number of qubits and their quality are steadily increasing. Another important factor is speed, measured in Circuit Layer Operations per Second (CLOPS). This number denotes how many quantum circuits can run on the quantum computer per time. It is one of the three important performance criteria of a quantum computer: scalability, quality, and speed.

In the foreseeable future, it should be possible to demonstrate the superiority of quantum computers for certain applications—that is, that they solve problems faster, more efficiently, and more precisely than a conventional computer. But building a powerful, error-corrected, general-purpose quantum computer will still take some time. Experts estimate that it will take at least another ten years. But the wait could be worth it. Like optical chips or new architectures for electronic computers, quantum computers could be the key to the mobility of the future.

In brief

When it comes to AI calculations, not only conventional microprocessors, but also graphics chips, are now reaching their limits. Companies and researchers worldwide are therefore working on new solutions. Chips in wafer format and light computers are close to becoming reality. In a few years, these could be supplemented by quantum computers for particularly demanding calculations.

Info

Text first published in the Porsche Engineering Magazine, issue 1/2022.

Author: Christian Meier

Image Sublines

Path: media/Images/img.png
Title: Cerebras' Wafer Scale Engine, 2022, Porsche AG
Subline: Giant chip: Cerebras' Wafer Scale Engine combines enormous computing power on a single integrated circuit with a side length of more than 20 centimeters.

Path: media/Images/img_1.png
Title: Lightmatter's Envise chip, 2022, Porsche AG
Subline: Computing with light: Lightmatter's Envise chip uses photons instead of electrons to calculate neural networks. The input and output data is supplied and received by conventional electronics.

Path: media/Images/img_2.png
Title: Heike Riel, Lead IBM Research Quantum Europe/Africa, 2022, Porsche AG
Subline: Heike Riel, Lead IBM Research Quantum Europe/Africa

Link Collection

Link to this article
https://newsroom.porsche.com/en_AU/2022/innovation/porsche-engineering-quantum-computers-computer-chip-optical-computers-28087.html

External Links
https://www.porscheengineering.com/peg/en/

Dr. Ing. h.c. F. Porsche Aktiengesellschaft
Porscheplatz 1
70435 Stuttgart