Booster for AI calculations

26/02/2020 Automated and autonomous driving functions are impossible to implement without AI. The required computing capacity is provided by special chips specialized in parallel computing. But researchers are also working on new, biologically inspired solutions as well as on quantum computers that promise even more computing capacity.

For decades, electronics have become increasingly prevalent in vehicles. Today, dozens of networked control devices control the engine, transmission, infotainment system and many other functions. Cars have long since become rolling computing centers—but now a new leap in computer power awaits them, because automated driving functions and autonomous driving require ever more powerful computers. And because the required performance cannot be achieved with conventional chips, the hour has come for graphics processors, tensor processing units (TPUs), and other hardware specially designed for the calculations of neural networks.

While conventional CPUs (central processing units) can be used universally, they lack the optimal architecture for AI. That is due to the typical calculations that occur during the training of and inference process with neural networks. “The matrix multiplications in neural networks are very elaborate,” explains Dr. Markus Götz of the Steinbuch Centre for Computing at the Karlsruhe Institute of Technology (KIT). “But these calculations are very amenable to parallelization—particularly with graphics cards. Whereas a high-end CPU with 24 cores and vector commands can perform 24 times 4 calculations per cycle; with a modern graphics card it's over 5,000.”

Graphics processors (GPUs, graphics processing units) are specialized for parallel work from the outset and have an internal architecture tailored for that purpose: GPUs contain hundreds or thousands of simple computation modules for integer and floating-point operations, which can simultaneously apply the same operation to different data (single instruction multiple data). They are therefore able to execute thousands of computing operations per clock cycle—for instance to compute the pixels of a virtual landscape or the matrix multiplications for neural networks. So it's no wonder that chips from the GPU manufacturer NVIDIA are currently ideally positioned as the workhorses for artificial intelligence in general and autonomous driving in particular. Volkswagen uses the US company's hardware, among others. “You need special hardware for autonomous driving,” says Ralf Bauer, Senior Manager Software Development at Porsche Engineering. “GPUs are the starting point; later, application-specific chips will presumably follow.”

NVIDIA currently offers the Xavier processes for autonomous driving specifically. A silicon chip is outfitted with eight conventional CPUs and one GPU specifically optimized for machine learning. For automated driving on level 2+ (limited longitudinal and lateral control with enhanced functionality based on standard sensors compared to level 2), the Drive AGX Xavier platform is available, which can execute a maximum of 30 trillion computing operations per second (30 TOPS, Tera Operations Per Second). For highly automated and autonomous driving, NVIDIA has the Drive AGX Pegasus (320 TOPS), under the control of which a test vehicle has driven as far as 80 kilometers without human

intervention through Silicon Valley. As the successor to Xavier, NVIDIA is now developing the Orin GPU, though little is currently known about its performance data.

Not all automobile manufacturers utilize GPUs. In 2016, Tesla began developing its own processors for neural networks. Instead of graphics processors from NVIDIA, the US-based company has been installing its FSD (Full Self-Driving) chip in its vehicles since early 2019. In addition to two neural processing units (NPUs) with 72 TOPS apiece, it also contains twelve conventional CPU cores for general calculations and a GPU for post-processing of image and video data. The NPUs, like GPUs, are specialized in parallel and thereby fast execution of addition and multiplication operations.

Google chip for AI applications

Google is a further newcomer in the chip business: since 2015, the technology company has been using self-developed TPUs in its data centers. The name comes from the mathematical term “tensor,” which encompasses vectors and matrices, among other elements. This is why Google's widely used software library for artificial intelligence is called TensorFlow—and the chips are optimized for them. In 2018, Google presented the third generation of its TPUs, which contain four “matrix multiplication units” and are said to be capable of 90 TFLOPS (Tera Floating Point Operations Per Second). The Google subsidiary Waymo uses TPUs to train neural networks for autonomous driving.

Application-specific chips like Tesla's FSD or the TPUs from Google only become economical at large unit numbers. One alternative is FPGAs (field-programmable gate arrays). These universally usable digital chips contain countless computing and memory blocks that can be combined with each other through programming and with which it is possible to essentially pour algorithms into hardware—like with an application-specific chip, but much more cheaply. FPGAs can be easily adapted to the specific requirements of an AI application (for instance specified data types), which yields benefits in terms of performance and energy consumption. The Munich-based start-up Kortiq has developed its AIScale architecture for FPGAs, which simplifies the neural networks for image recognition and so optimizes the calculations that the requirements on the hardware diminish significantly and results are available up to ten times faster.

Some researchers are pursuing an even closer relationship to the functioning of nerve cells for AI-specific chips. Researchers at Heidelberg University have developed the neuromorphic system BrainScaleS, whose artificial neurons are implemented as analog switches on silicon chips: the cell body consists of some 1,000 transistors and two capacitors, with the synapses requiring roughly 150 transistors. Individual cell bodies can be combined as modules to form various types of artificial neurons. These synapses can, as in nature, form strong connections, and there are also excitatory and inhibitory types. The output of the neurons consists of “spikes,” short voltage pulses lasting a few microseconds that function as inputs for the other artificial neurons.

Energy-efficient neuro-chips

But BrainScaleS is not just used to research the human brain. The technical neurons can also be used to solve technical problems—such as object detection for autonomous driving. On the one hand, they offer high computing capacity of approximately a quadrillion computing operations (1,000 TOPS) per module with 200,000 neurons. On the other hand, the analog solution also uses very little energy. “In digital circuits, for example, there are some 10,000 transistors used for each operation,” explains Johannes Schemmel of Heidelberg University. “We get by with substantially fewer, which enables us to achieve roughly 100 TOPS per watt.” The researchers have just developed the second generation of their circuits and are talking to industry partners about possible collaborations.

Quantum power from the cloud

In the future, even quantum computers could be used in the field of AI. Their fundamental unit is not the binary bit, but the qubit, with an infinite number of possible values. Thanks to the laws of quantum mechanics, calculations can be highly parallelized and thereby accelerated. At the same time, quantum computers are difficult to implement because qubits are represented by sensitive physical systems like electrons, photons, and ions. This was demonstrated, for example, with the IBM Q System One, which the company introduced at the CES 2019 electronics trade show in Las Vegas. The interior of the quantum computer must be fastidiously shielded against vibrations, electrical fields, and temperature fluctuations.

Nerve cells and artificial neurons

Nerve cells receive their signals from other neurons via synapses that are located either on the dendrites or directly on the cell body. Synapses can have either an excitatory or inhibitory effect. All inputs are totaled at the axon hillock and if a threshold is exceeded in the process, the nerve cell fires off a roughly millisecond-long signal that propagates along the axon and reaches other neurons.

Artificial neurons mimic this behavior more or less exactly. In conventional neural networks with multiple layers, each “nerve cell” receives a weighted sum as an input. It consists of the outputs of the neurons of the preceding layer and the weighting factor wi, in which the learning experience of the neural network is stored. These weighting factors correspond to the synapses and can also be excitatory or inhibitory. A configurable threshold value determines, like in a nerve cell, when the artificial neuron fires.

Learning from and inference with neural networks

Natural and artificial neural networks learn from changes in the strength of synaptic connections and

the weighting factors. In deep neural networks, during training, data is fed to the inputs and the output compared with a desired result. Using mathematical methods, the weighting factor wij is continually readjusted until the neural network can reliably place images, for example, in specified categories. With inference, data is fed to the input and the output is used to make decisions, for example.

In both training and inference in deep neural networks (networks with multiple layers of artificial neurons), the same mathematical operations occur repeatedly. If one sums both the outputs of the neurons of layer 1 and the inputs of the neurons of layer 2 as column vectors, all calculations can be represented as matrix multiplications. In the process, numerous mutually independent multiplications and additions occur that can be executed in parallel. Conventional CPUs are not designed for that—and that is why graphics processors, TPUs, and other AI accelerators are vastly superior to them.

In brief

Conventional computer chips reach their limits when it comes to calculations for neural networks. Graphics processors and special hardware for AI developed by companies such as NVIDIA and Google are much more powerful. Neuromorphic chips are substantially similar to real neurons and work very efficiently. Quantum computers could also boost computing capacity enormously.

Info

Text: Christian Buck
Contributors: Ralf Bauer, Dr. Christian Koelen

Text first published in the Porsche Engineering Magazine, issue 02/2019

Link Collection

Link to this article
https://newsroom.porsche.com/en_US/technology/porsche-engineering-booster-ai-calculations-autonomous-driving-19957.html

Media Package
https://pmdb.porsche.de/newsroomzips/1232d5fa-9cad-4b09-a92b-8883b656fb86.zip

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