Rising to the acceleration challenge

The AI hardware acceleration challenge will most likely be solved by using a different processing model, according to **Rob Taylor**.

Artificial Intelligence (AI) and deep learning is the future of computing. Intelligent machines that understand the world as humans do, interpret our languages and learn from data will habitually be used to resolve problems too complex for the human brain.

But progress in AI is being blocked by the simple lack of computing power. While scientists are building advanced algorithms, they do not yet have the hardware necessary to train machines on these algorithms nor for machines to execute the algorithms and apply their learning to new data. Although DeepMind's AlphaGo algorithm famously outsmarted the world's best Go player last year, it reportedly required 1202 CPUs and 176 GPUs (or in fact, Google's own TPUs) to do so – not exactly practical.

New solutions to hardware acceleration are required, so what is it about AI that requires a new approach to computing power and what are the possible options?

Basic forms of AI can function on traditional processors; for instance, IBM's Watson runs on a combination of power processors, GPUs and CPUs. Although Watson is often held up as the pinnacle of AI development, its functionality is limited to finding patterns and insights hidden in data. IBM is already looking to far more powerful processors to provide answers to questions where there is insufficient data to find patterns and the number of potential permutations is too vast to be processed by classical computers.

Deep learning is a new software model that requires a different type of computer platform. In deep neural networks, algorithms learn from data and examples, but effectively write their own software. This means software-neurons and connections must be trained in parallel, rather than sequentially.

Advances in CPUs have slowed and the marginal gains being delivered will not be sufficient to run deep neural networks effectively. A processing model is needed that can execute programmer-coded commands and the parallel training of deep neural networks.

Three approaches have emerged to meet the hardware acceleration challenge – GPUs, TPUs and FPGAs.

"GPUs may have been instrumental in kickstarting the AI revolution, but FPGAS will be the driver of future developments"
Rob Taylor

Graphics processing

GPUs have, so far, been the preferred alternative to CPUs for training deep neural networks because their design enables them to handle large quantities of basic computations in parallel. Originally designed to draw real-time images quickly, such as video game graphics, by rendering at the same time the millions of lines of code that make a complete picture, GPUs are designed for specific types of mathematical operations such as matrix multiplications. These operations are also used in the training of deep neural networks and are the most computationally intensive.

GPU-accelerated computing – described by GPU market-leader Nvidia as 'a new computing model that uses massively parallel graphics processors to accelerate applications

Adobestock

also parallel in nature' – has been instrumental in triggering and driving the advance of AI and deep learning so far, but it does have restrictions.

The main limitation is that GPUs are designed predominantly for graphics and not deep learning, where data needs to be structured in a specific format. This can be particularly challenging when dealing with complex data sets such as speech or video. GPUs also have limited reprogrammability, which is a concern given the rapid evolution and development of AI technologies. Finally, GPUs do not currently provide lightning-fast connections between parallel onboard computing cores, nor instant access to the memory required to store complex models and data, so latency remains an issue.

Tensor processing unit

Last year, Google announced it had been using an internally developed chip – the TPU – for more than a year to accelerate deep learning applications and that it had been used in AlphaGo's victory. Google claims its TPU – now available in Google Cloud – can deliver up to 30 times more performance and up to 80 times more performance/W than CPUs and GPUs.

While this increase in performance is impressive, the TPU is inherently limited because it can only be used within Google environments and with Google TensorFlow software. Developing an ASIC for a specific use – such as deep learning – only makes sense for a dominant player such as Google. Smaller businesses and start-ups looking to incorporate deep learning into their own products and business will need a far more flexible solution.

FPGAs

FPGAs provide a third choice. Because they contain thousands of reprogrammable logic blocks that can perform various processes at the same time, FPGAs offer massive concurrency and are ideal for parallel, low-latency, high throughput processing. As with GPUs and TPUs, FPGAs speed processing times – up to 100 times faster than the same code running on traditional CPUs.

But their main advantage over GPUs and TPUs is the ability to reconfigure the hardware after deployment, as and when algorithms evolve, increasing agility and flexibility. There are also significant efficiency gains, with the potential for tenfold improvements in power usage.

Author profile:
Rob Taylor is CEO Reconfigure.io

Recent developments, including Amazon EC2 F1 Instances, provide the ability to program FPGAs remotely in the cloud using high-level languages such as Go, making FPGAs far more accessible and cost effective.

Microsoft recently unveiled Project Brainwave, a deep learning acceleration platform designed for real-time AI that allows deep neural networks to be mapped to a pool of remote FPGAs and called by a server, reducing latency and increasing throughput. Project Brainwave will eventually be available in Azure, allowing direct access to the AI system through the Microsoft cloud. As FPGA usage continues to become cheaper and more realistic, adoption will become more widespread.

While GPUs may have been instrumental in kickstarting the AI revolution, it will most likely be FPGAs that will drive future developments. While other alternatives may deliver the increased speeds and parallel computing necessary for deep learning, the flexibility, reconfigurability, accessibility and efficiency of FPGAs make them a sustainable solution to the AI hardware acceleration challenge.

Neuromorphic computing

The processing capabilities of FPGAs have been demonstrated by BrainChip, a developer of software and hardware accelerated solutions for AI, with the release of the BrainChip Accelerator, an acceleration board which uses FPGAs to process visual data.

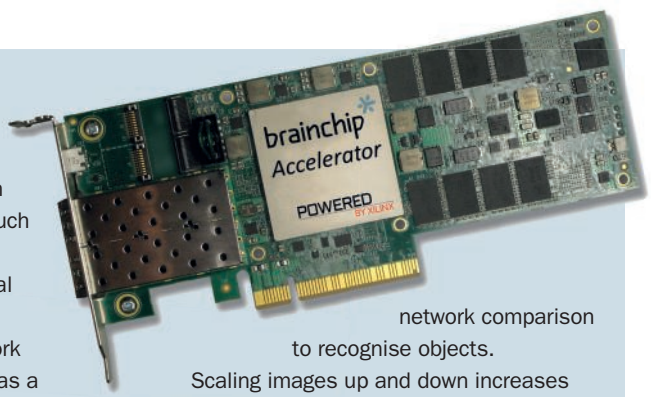
The Accelerator is an eight lane, PCI-Express add-in card that increases the speed and accuracy of object recognition in BrainChip's Studio software, while increasing the simultaneous video channels of a system to 16 per card.

The low-power card is said to be easy to install within existing video surveillance systems without the need to upgrade power systems or thermal management.

Studio software is used to process large amounts of archived or live streaming video and the Accelerator add-in card mean more video can be processed much faster.

Said to be the first commercial implementation of a hardware-accelerated spiking neural network system, the Accelerator is seen as a significant milestone in the development of neuromorphic computing, a branch of artificial intelligence that simulates neuron functions.

Processing is done by six BrainChip Accelerator cores in a Xilinx Kintex Ultrascale FPGA. Each core performs user-defined image scaling, spike generation, and spiking neural network comparison to recognise objects.



Scaling images up and down increases the probability of finding objects, and due to the low-power characteristics of spiking neural networks, each core consumes approximately 1W while processing up to 100frame/s.

In comparison to GPU-accelerated deep learning classification neural networks, this represents a 7x improvement of frame/s/W.