Illustration: illumate/stock.adobe.com

The cost of learning

How are embedded system designers addressing the incredible complexity associated with machine learning? By **Chris Edwards**

Machine learning has been through numerous slumps, each following a spasm of technological over-enthusiasm. But it may be the only way to deal with the incredible complexity of situations that embedded-systems designers now face – something that is now being reflected in the strategies of embedded-processor designers.

Rene Haas, IP products group lead at ARM, says: "We believe machine learning is one of the most significant change that is hitting our computing landscape. We believe that years from now people will not be looking at machine learning as a unique category where computers learn but rather it will be native to everything that computers do."

Although people tend to associate machine learning with image and audio recognition through services like Apple's Siri, the systems already in the market send a lot of data to the cloud to be processed. That is changing rapidly, says Haas.

"What is rising is the amount of inference and training that can take place at the edge. The level of analytics, the level of learning, level of sophistication performed locally has moved much faster than I think everyone has anticipated," Haas explains, pointing to bandwidth and server-capacity problems as chief concerns.

"Why not send the data to the cloud? There isn't enough bandwidth in the world; the internet would simply collapse. Google themselves said if everyone in the world with an Android phone were to use their voice assistant for just 3 minutes they would have to double the number of servers."

"The level of analytics, of learning and of sophistication performed locally has moved much faster than I think everyone has anticipated." Rene Haas

Beyond the consumer-visible applications of voice response and image recognition, machine learning is moving into deeply embedded applications. Ceva decided to incorporate machine learning into its Penta-5G IP cores destined for the next generation of mobile phones. Emmanuel Gresset, director of business development at Ceva, says the increases in bandwidth promised by 5G are made possible by the combination of techniques such as MIMO and beamforming: "This is efficient only if link adaptation is very well done. But it's a complex task. If the link adaptation is not accurate you will not achieve the throughput and you will consume a lot more power.

"We looked at traditional, more algorithmic approaches. The [parameter] dimensions are so large that it would require a lot of memory to do that. The complexity increases exponentially. That's why we came up with a neural network-based approach. The AI processor takes as input the channel conditions and it computes the best transmit parameters," Gresset notes.

Trade-offs

The problem that faces designers of machine-learning accelerators is the trade-off between flexibility and energy consumption. There are many implementations, and computer scientists are rapidly

adding modifications that attempt to cater for different types of data and application. There is no guarantee that currently fashionable convolutional neural networks (CNNs) will remain dominant as researchers work on other options ranging from techniques based on statistical processing such as Gaussian processes to architectures that borrow even more from neurology than CNNs.

General-purpose processors offer the ability to handle the widest range of options but they have seen their best days, argues Steve Mensor, vice president of marketing at Achronix: "As the tasks have become more complex, clock speed have maxed out, meaning we need more CPUs. The only thing that keeps going up and to the right is transistor count. All the other functions have tapered off. So, the CPU's ability to keep up with market requirements appears to have been lost. Accelerators will be the brawn. A lot of start-up companies are looking at different architectures but they are all working on highly parallelised compute units that are targeted for a particular application."

To maintain a balance between efficiency and programmability, Ceva, Cadence Design Systems and others favour augmenting parallelised DSPs with specialised instructions working alongside more specific hardware accelerators. The accelerators home in on frequently used operations such as the thresholding step that typically happens after a neuron has incorporated data from all its inputs.

Other vendors are building AI accelerators into programmable-logic fabrics to take advantage of the way that these devices make it easy to route data between processing elements without the energy-sapping need to shift data to and from memory at each stage. Having seen its FPGA, as well as those from chief competitor Intel PSG, move into data-centre blades for various acceleration tasks, including machine learning, Xilinx aims to capitalise on the shift with its forthcoming "Project Everest" devices. These will link the core programmable-logic fabric to more specialised programmable accelerators using an onchip network that CEO Victor Peng says he expects will be reconfigured on the fly.

Mensor argues the embedded FPGA technology has advantages over the off-the-shelf variety. Embeddable IP provides the ability for customers to add their own optimisations to the programmable fabric. He says the company has worked with a customer on building specialised cores for CNNs: "They handle different types of object-recognition applications and they use an array of kernels. We have cores where we have changed the DSP to the requirements they have, which cuts 50 per cent of the silicon area of those blocks. Overall, we get a 35 per cent die-size reduction on the final device."

Peng says embedded-FPGAs companies have the disadvantage of only having access to a programmable-logic technology and not the various other IP cores the standalone-device makers can wrap around them, such as high-speed Serdes I/O ports.

Mensor says the I/O itself becomes less important because the embedded-FPGA technology makes it easier to squeeze more onto a single die. He points to the existing data-centre applications where, according to his company's estimates, as much as 60 per cent of the FPGA is dedicated to I/O. Very wide local-memory interconnect cuts the delay and power to reach locally stored data.

"People are realising there are limits to the amount of memory you can put onchip. So, the question is: can we find other methods?" Pulin Desai

"The CPU's ability to keep up with market requirements appears to have been lost. Accelerators will be the brawn." Steve Mensor

However, CNNs have prodigious appetites for data that can make it difficult to bring everything onchip. Pulin Desai, director of AI and imaging IP product development at Cadence, says: "People are realising there are limits to the amount of memory you can put onchip. So the question is: can we find other methods?"

A number of the vendors are working on techniques to reduce the amount of data a trained CNN needs to access. Most have already implemented narrow datapaths on the basis that high data resolution is important in training – which can be offloaded to remote servers – but for the real-time inferencing processes, bit widths as narrow as 8bit offer reasonable performance. There are indications that some of the weight calculations performed by artificial neurons can scale down to binary. Often there are redundant paths that can be pruned away entirely. Another option is to compress data at rest and only decompress it when it is needed.

"There is a lot of activity in the compression of weights as well as pruning and decompressing on the fly," Desai notes.

Ultimately, the memory issue in machine learning and other data-intensive algorithms that look to dominate computing in the decade ahead may force a more radical change in architecture. Server designers are already pushing more processing into flash-storage subsystems. Although the programming will be much more difficult, processors distributed through memory may be the only way to deal with the need to boost performance without suffering a massive increase in energy consumption.