Why AI could cause 'full fat' computing to interrupt the mobile revolution

Originally published at https://www.intelligentautomation.network/artificial-intelligence/articles/why-ai-could-cause-full-fat-computing-to-interrupt April 20, 2021

The launch of AMD's Ryzen family of CPUs this month represents an early bid to engage the Machine Learning market at a consumer level. But is it approaching AI from the wrong direction?

Photo by Gabriel Santiago on Unsplash

On the surface, the launch of AMD's Ryzen processor family is just another entertaining show-down in the venerable CPU cold war. AMD long ago lost the high ground to Intel, when its own price-cutting and market positioning forced the public (particularly the early PC gaming community) to choose between economy and processing grunt. The public went for the higher-priced power in the end, and AMD languished as the 'Pepsi' of the CPU challenge for over a decade.

This time AMD is back with what can be interpreted either as a vanguard gesture towards the future of AI-inclusive consumer hardware, or just a gimmick - because the Ryzen processor class contains hardware-level protocols dedicated to Machine Learning.

Ryzen CPU iterations take granular sales tactics to an almost unprecedented level, with six offerings in various stages of roll-out, potency and pricing. But underlying them all is a 'Smart Prediction' feature touted as being able to optimize the path of Machine Learning/Neural Network instructions through the processor.

Local help

Zen's Machine Learning facility appears to be based around Perceptron branch predictors. The Perceptron caused much excitement during the first AI revolution of the 1960s, adding a framework for layers and higher-level/resolution results from machine-driven Neural Network processes.

TAGE and OGHEL branch predictors have picked up the mantle in the nearly sixty years since the Perceptron's debut, and these do not appear to be directly addressed in Ryzen, as far as early indications can tell.

So Ryzen's hardware AI component could be perfunctory dressing seeking a market edge against Intel's Core i9 offering. It's not likely to put Intel on the ropes, but does open up the debate about whether the next evolution in computing should be entirely about aggregate processing and the open source/widely-contributive model, or whether it's time for consumer-level hardware to get its hand in its pockets and start making a more active contribution to AI development.

Bloat vs distribution

There have been some portents that consumer-level hardware might engage with AI directly in the next five years. Though the NASA/military development pipeline has largely been replaced by the free market in the last thirty years, the private sector has nonetheless anticipated AMD's innovation with - amongst other offerings - IBM's data center server blades aimed at the AI end of the data center sector.

Intel instead appears to see the future of consumer-level Machine Learning as driven by the open source community, in libraries such as Caffe, Theano, Apache Spark and its own Trusted Analytics Platform (TAP). Assuming the company is not about to announce that i10 will be hard-wired for neural networks, Intel appears to be showing a more mainstream and hardware-agnostic take on how AI and Machine Learning is likely to diffuse to the consumer space.

However it's interesting to note that Intel has taken down a PDF that revealed in 2016 that dedicated Deep Learning instructions were planned for its AVX-512 family of CPU instructions, though the document can be found elsewhere .

> ***In any case, it's possible that an increasingly essential new technology (neural networks) is shortly going to vie for resident space in a consumer market which has become more mobile - and***

necessarily minimalist - than the geeks of 1999/2000 could possibly have dreamed.

When this happens, you traditionally end up with initially bloated and heavy hardware which then undergoes a gradual process of optimization and improvement over a period of around 10-15 years. It would be a movement in complete opposition to the multi-billion dollar trend currently dominating the mobile space.

It's not inevitable. The alternate paradigm positions individual mobile devices as network nodes, relatively dumb terminals contributing to aggregate computing resources in the same way that individual circuits or cores contribute to the net performance of a multi-core CPU.

'Toy Story' prefigures the AI revolution?

But in that case, the entire problem becomes one of network latency, and one sacrifices physical heaviness for network heaviness - responsive lag. In that model, the successful advent of reliable pan-continental 5G could end up being a critical component in placing live neural network responses inside the 'pedestrian' experience of affordable technology.

Perhaps the closest technological analogy of the last thirty years, in terms of harvesting and systematising a new and rabid need for high compute cycles, has been the evolution of network and local rendering systems for Computer Generated Imagery (CGI).

The seminal *Toy Story* (1995), at the time the most ambitious all-CGI project ever attempted, required so much processing power that even the secretaries at Pixar would find a custom-made network rendering node had taken over their PCs after a brief bathroom break, in an effort to render another 1/5th of Woody's latest frown.

This led to the advent of the 'render-farm' as a 1990s industry staple: banks of CPUs networked together to form unified processing powerhouses that could blaze through polygons and shaders - at least, relative to the standard of the time.

This evolved into custom commercial hardware and services which brought all that capacity into a box, inevitably leading to 'Renderfarm renderfarms', which networked clusters of these consolidated machines together to raise capacity or throughput (not usually both).

Now, after the cloud and remote processing revolution of the last ten years, the huge proliferation of global visual effects companies routinely buy render computing cycles off large-scale vendors such as Amazon Web Services in order to fulfil increasing demand for CGI.

The power behind machine learning

There are other budgetary concerns likely to influence the direction in which mainstream neural networks will develop. One is electricity consumption - at its current semi-conceptual stage of evolution, Machine Learning tends towards being power hungry and resource intensive, depending on the level of resolution required.

Perhaps the greatest field test of distributed computing in the GPU era of data analysis was the PlayStation 3's contribution to Stanford University's Folding@home distributed computing project, which Sony removed from the console's capability in 2012, to general bewilderment. The company announced that the PS3 participation ended after 'discussions with Stanford University', though there was speculation about the effect the idle-process folding system might have both on console component longevity and power consumption.

One commenter claimed that folding proteins all day on his PS3 raised his electricity bills by $216 a year.

Even distributed computing systems which are ruminative in nature (i.e. projects such as SETI where latency is not an issue) also have a net effect on bandwidth consumption in a commercial environment where users may be data-capped - an argument in favour of having more local processing power sending fewer packets that are more 'processed'; but again, at the cost of heat, lifecycle erosion, power consumption and responsiveness in portable devices which have been relying on single-process thread models for a long time now.

Considering the possible range of trade-offs, the benefits of 'local' neural networks and Machine Learning hardware components will have to be made unequivocally compelling in order to provide impetus either for the potential infrastructural demands or changes to the local device capacities to be met by the market. It's a commercial singularity that doesn't appear to have happened yet.

And that's assuming that another bulky prototype hardware platform doesn't muddy the issue even further in anything like the near future.#