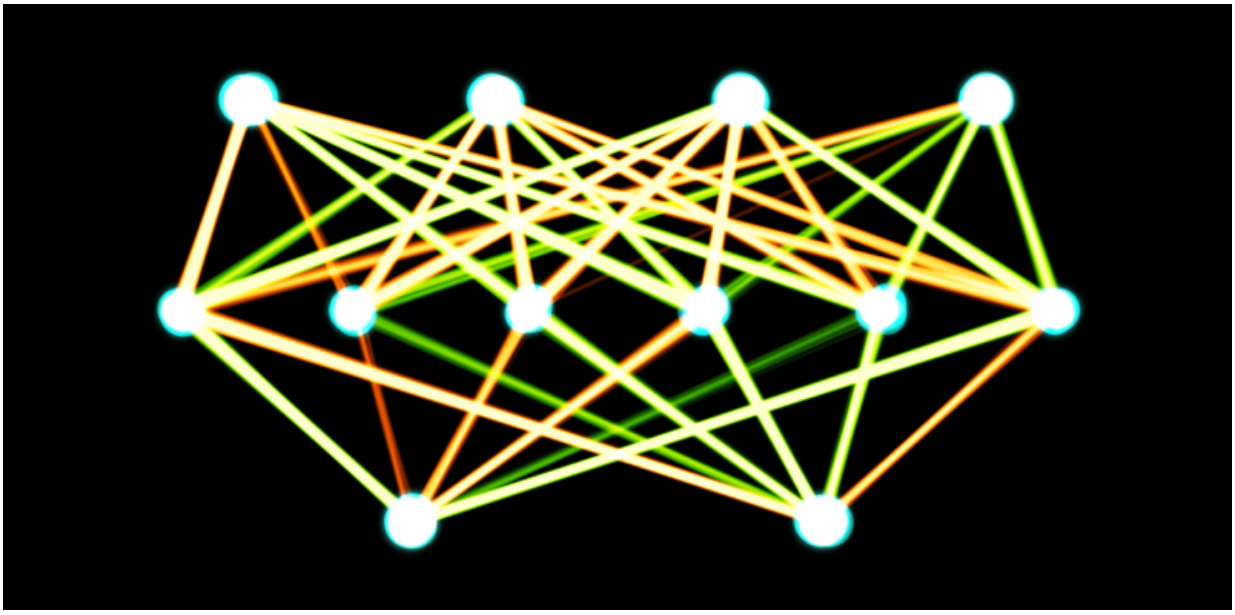
What powers Facebook and Google's AI – and how computers could mimic brains

January 6 2016, by Thomas Nowotny, University Of Sussex



Credit: Akritasa, CC BY-SA

Google and Facebook have open sourced the designs for the computing hardware that powers the artificial intelligence logic used in their products. These intelligent algorithms power Google's search and recommendation functions, Facebook's Messenger digital assistant, M – and of course both firms' use of targeted advertising.

Facebook's bespoke computer servers, codenamed Big Sur, are packed

with graphics processing units (GPU) – the graphics cards used in PCs to play the latest videogames with 3D graphics. So too is the hardware that powers Google's TensorFlow AI. So why is artificial intelligence computing built from graphics processors instead of mainstream computer processors?

Originally GPUs were designed as co-processors that operated alongside a computer's main central processing unit (CPU) in order to off-load demanding computational graphics tasks. Rendering 3D graphics scenes is what is known as an embarassingly parallel task. With no connection or interdependence between one area of an image and another, the job can be easily broken down into separate tasks which can be processed concurrently in parallel – that is, at the same time, so completing the job far more quickly.

It's this parallelism that has led GPU manufacturers to put their hardware to a radically different use. By optimising them so that they can achieve maximum computational throughput only on massively parallel tasks, GPUs can be turned into specialised processors that can run any parallelised code, not just graphical tasks. CPUs on the other hand are optimised to be faster at handling single-threaded (non-parallel) tasks, because most general purpose software is still single-threaded.

NVIDIA Tesla M40 GPU Accelerator. Credit: NVIDIA news

In contrast to CPUs with one, two, four or eight processing cores, modern GPUs have thousands: the NVIDIA Tesla M40 used in Facebook's servers has 3,072 so-called CUDA cores, for example. However, this massive parallelism comes at a price: software has to be specifically written to take advantage of it, and GPUs are hard to program.

What makes GPUs suitable for AI?

One of the reasons GPUs have emerged as the supercomputing hardware of choice is that some of the most demanding computational problems happen to be well-suited to parallel execution.

Facebook Big Sur server containing 8 NVIDIA Tesla M40 GPUs. Credit: Facebook

A prime example is deep learning, one of the leading edge developments

in AI. The neural network concept that underpins this powerful approach – large meshes of highly interconnected nodes – is the same that was written-off as a failure in the 1990s. But now that technology allows us to build much larger and deeper neural networks this approach achieves radically improved results. These neural networks power the speech recognition software, language translation, and semantic search facilities that Google, Facebook and many apps use today.

Training a neural network so that it "learns" works similarly to establishing connections between neurons and strengthening those connections in the brain. Computationally, this learning process can be parallelised, so it can be accelerated using GPU hardware. This machine learning requires examples to learn from, and this also lends itself to easy acceleration using parallel processing. With open source machine learning tools such as the Torch code library and GPU-packed servers, neural network training can be achieved many times faster on GPU than CPU-based systems.

Titan supercomputer at the Oak Ridge National Laboratory. Credit: Oak Ridge National Laboratory

Are GPUs the future of computing?

For decades we have become accustomed to the version of Moore's law which holds that computer processing power will roughly double every two years. This has mainly been achieved through miniaturisation, which leads to less heat generation, which allows CPUs to be run faster. However, this "free lunch" has come to an end as semiconductors have been miniaturised close to silicon's theoretical, elemental limits. Now, the only credible route to greater speeds is through greater parallelism, as demonstrated with the rise of multi-core CPUs over the last ten years. GPUs, however, have a head start.

Besides AI, GPUs are also used for simulations of fluid and aerodynamics, physics engines and brain simulations, to name just a few examples. Some of the world's most powerful computers, such as the Titan supercomputer at Oak Ridge National Laboratory, currently the world's second fastest supercomputer, are built on Nvidia's GPU accelerators, while competitors include Intel's Phi parallel co-processor that powers Tianhe-2, the world's fastest supercomputer. However, not all problems are easily parallelisable, and programming for these environments is difficult.

Arguably, the future of computing, at least for AI, may lie in the even more radically different neuromorphic computers. IBM's True North chip is one, with another under development by the €1 billion Human Brain Project. In this model, rather than simulating neural networks with a network of many processors, the chip *is* the neural network: the individual silicon transistors on the chip form circuits that process and communicate via electrical signals – not dissimilar to neurons in biological brains.

Proponents argue that these systems will help us to finally scale up our neural networks to the size and complexity of the human brain, bringing

AI to the point where it can rival human intelligence. Others, particularly brain researchers, are more cautious – there may well be a lot more to the human brain than just its high number and density of neurons.

Either way, it's likely that what we now learn about the brain will be through the very supercomputers that are designed to ape the way it works.

This article was originally published on The Conversation. *Read the* original article.

Source: The Conversation

Citation: What powers Facebook and Google's AI – and how computers could mimic brains (2016, January 6) retrieved 19 April 2024 from https://phys.org/news/2016-01-powers-facebook-google-ai-mimic.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.