MONASH University Engineering

Department of Electrical and Computer Systems Engineering

ECE4094/5 Final Year Project 2020

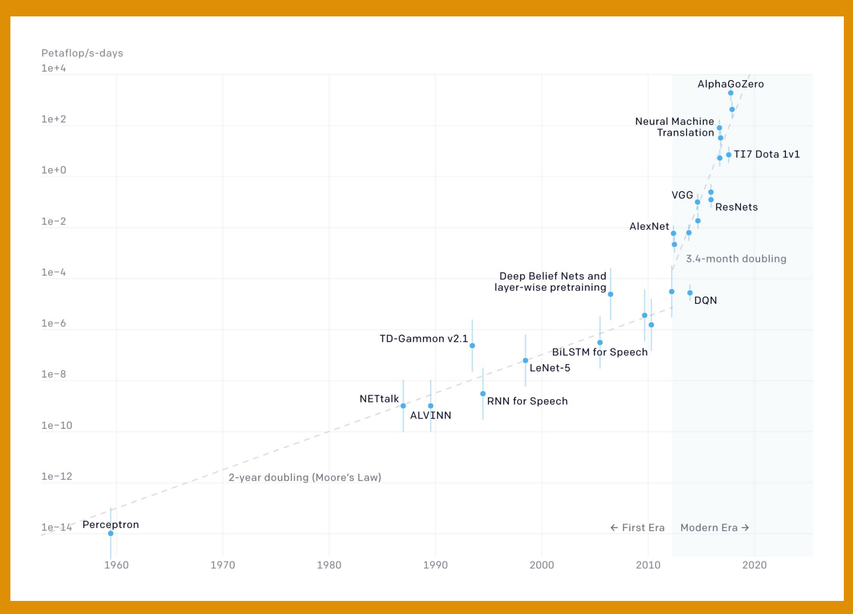
Nicholas Greig

Supervisor: Dr Lowery

On-chip Photonics for Feedforward Optical Neural Network Accelerators

The problem:

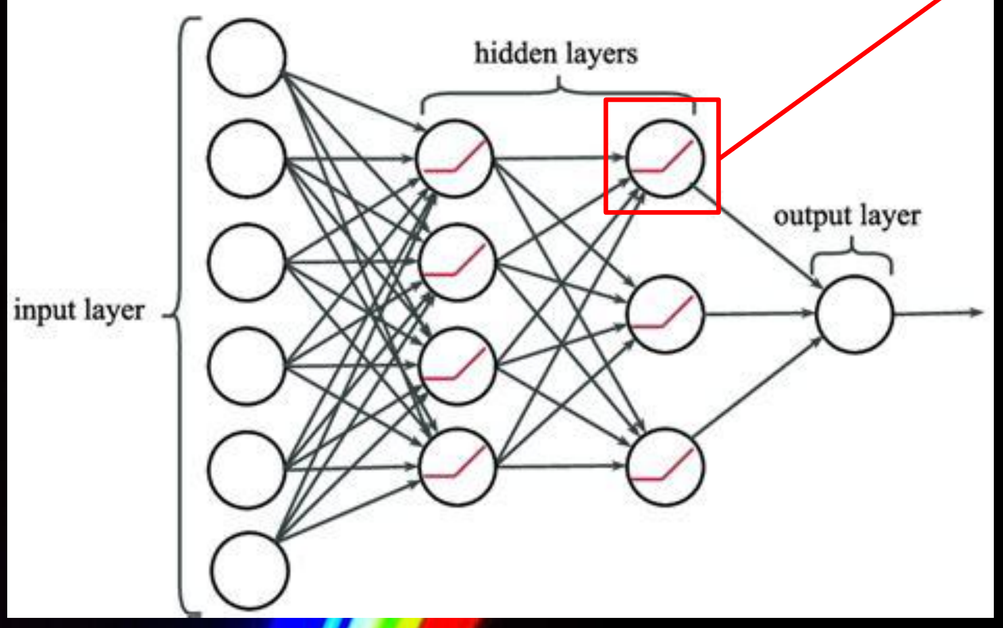
The amount of computation required to train state of the art deep learning models has doubled every three or so months since 2012. During this time, the efficiency of training deep learning models has doubled every 2 or so years, all whilst Moore's Law is winding down. This mismatch means that the energy, time, resources and money required to train a relevant deep learning model far outstrips Moore's Law, and is untenable in the long run. As an example of the magnitude of the current expenses required to train a large language model, training OpenAI's GPT-3 in the cloud would cost somewhere between 2 and 12 million USD from start to finish

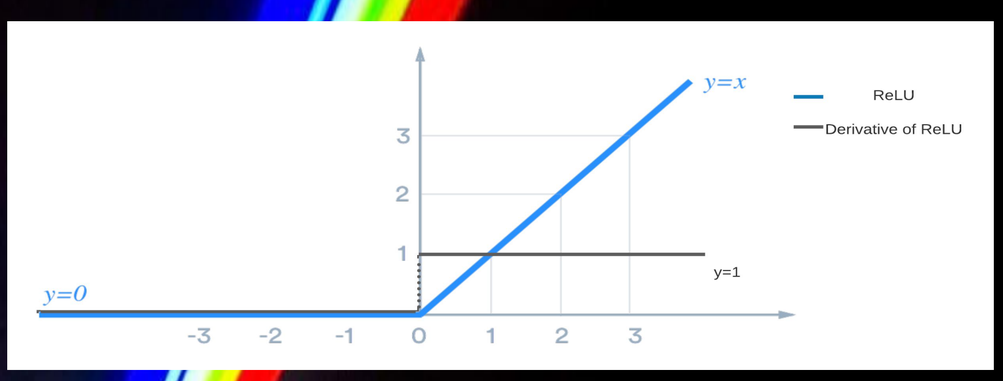


ReLU function

The rectified linear unit (ReLU) function is used in neural networks as a nonlinear activation function that allows deep neural networks to become universal function approximators, meaning they can be used to model any describable function. The ReLU function has seen widespread use as it is efficient to calculate on digital computers, suitably expressive for neural networks, and is inspired by biological neurons.

A multilayer neural network with ReLU activation function





Project Aim

The aim of this project was to assess the viability of using neuromorphic photonic hardware as a substrate for neural network computation.
Photonics is the science of generating, manipulating and detecting photons.
Neuromorphic computing is the art and science of computing using physical analogs to biological neurons. Colocating memory and processing power on neuromorphic hardware allows for high bandwidth, low energy consumption calculations to take place.
Neuromorphic photonics is the use of optical signals to help approximate the functions of biological neurons.

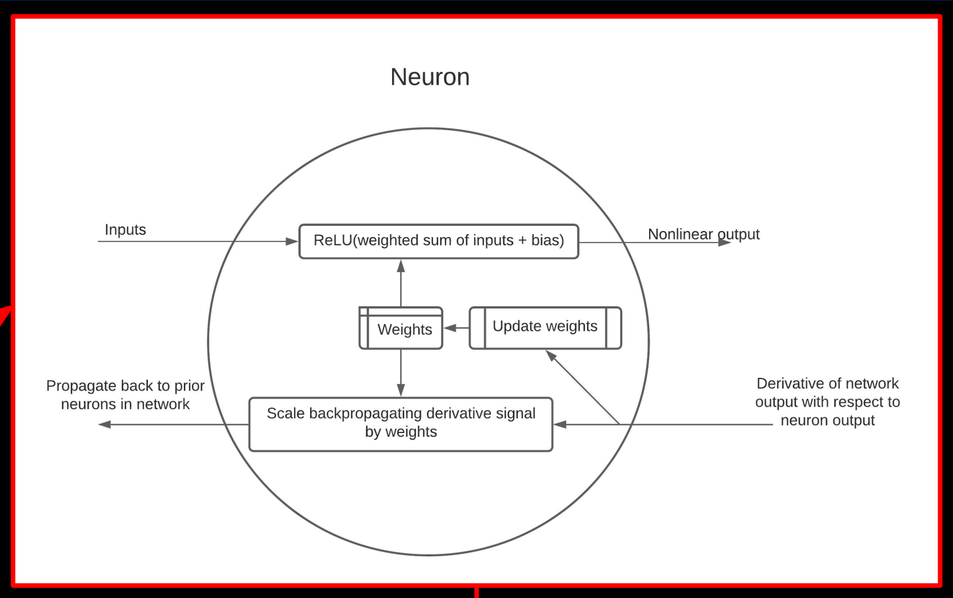
Why is neuromorphic photonics a good fit for deep learning workloads?

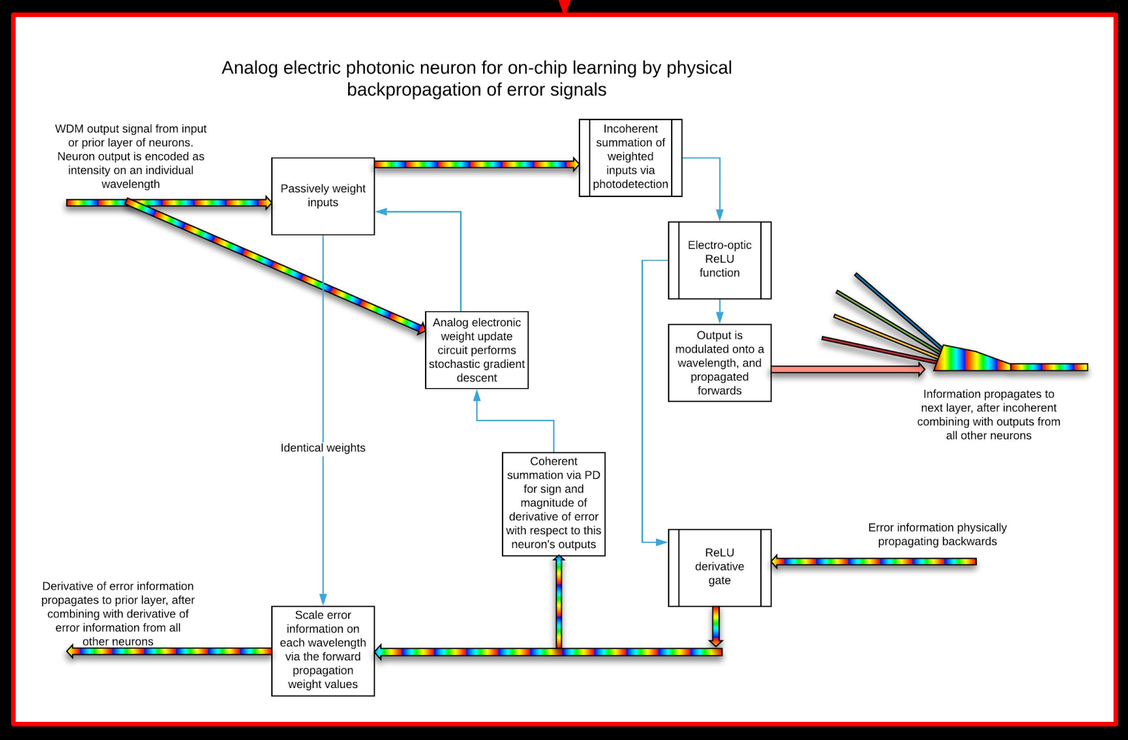
The matrix multiplication is the workhorse behind deep learning workloads. Matrix multiplications can be boiled down, at their core, to multiplications and additions.

In the photonic domain, additions can be performed passively, by interfering two coherent wavelengths together constructively. Scalar multiplication of a signal by any value between zero and one can be performed passively using microring resonators with diameters of <10μm.

Project outcomes:

During this project, a monolithically integratable design for a photonic ReLU neuron with the capacity for on-chip learning was developed. Numerical correctness of the SGD weight update circuit designed for learning was verified by simulation in Simulink.





The theoretical results:

For a 10-layer neural network, provided the entire network can fit on the chip, and the surrounding electronics can operate fast enough, the proposed architecture can train on 1.6 billion samples per second. That is not a typo. 1.6 billion samples per second means you could train on each of the 14 million images in the Imagenet database 115 times per second.