Photonic Chips for Deep Learning Neural Networks: Advancements in Performance Speed and Efficiency

Shannon Hull

Problem Statement:

In 1965, American engineer Gordon Moore predicted that each year, the number of transistors that could be put onto a single silicon chip would double. The prediction, which at one point actually became an underestimate, was considered by some circles to be so steadfast that it was referred to as Moore's Law.[12] These silicon chips have been the basis for digital computing for decades. Impressive advancements in computing, which now rival the performance of the human brain,[1] have been possible via multi-layered neural networks. These networks allow for a specific type of machine learning known as "deep" learning.[2]

The rapid pace at which the demand for computing power grows, however, has come at a price. Even with accessibility to graphical processing units (GPU's) for non graphical purposes, the race to advance deep learning often outpaces current technology.[1] Moore's Law is no longer accurate to real-life conditions,[1,3] and the work that is being done to improve traditional digital computing has resulted in an inefficient use of both energy and resources.[3] These setbacks have led scientists to seek out alternatives to digital deep learning networks that produce results comparable to current technology's, and are more efficient.

The answer to this call may lie in light - that is, the use of photons to perform calculations in deep learning neural networks. Although modern photonic technology may be novel, optical computing itself is not. In the mid -1900s, analog computers using light were able to process synthetic-aperture radar (SAR) image data. [3,4] At this point in history, digital computers were not yet powerful enough for this level of mathematical calculating.[3,4] The claims made by Moore's Law shortly after this pushed the focus on analog computing aside, but things are now coming full circle.[3]

Basic Background on Technology:

In a general sense, deep-learning neural networks are composed of artificial "neurons" that take a certain input, which is acted upon by a mathematical function and produces an output that subsequently becomes the input for the next connected neuron.[1] The connections run on a series of linear algebra calculations, specifically multiply-and-accumulate operations, based on mathematical arrays called matrices.[1] Light obeys Maxwell's equations, a group of equations that serve as the basis for electromagnetism, optics and circuits. These are linear equations, meaning in light-based networks, the outputs will be proportional to the inputs.[1] This restriction can be taken advantage of for the purposes of deep learning. This paper will focus on the ways in which photonic technology is being used to address the challenge of combining high performance speed with reduced energy consumption in neural networks, while also addressing the current limitations of this technology.

Current Advancements and Applications:

Several companies have dedicated research to the development and optimization of photonic computer chips. Lightmatter, a Boston-based computer hardware manufacturer specializing in photonic technology, has made significant progress in developing their own chips. They presented its "Mars" chip at an August 2020 Hot Chips conference.[4] This chip uses a component called the Mach- Zehnder Interferometer (MZI), one that's been shrunk to a size that allows several to be integrated together on one chip.[4] These devices split an input beam of light in two. The path that each beam takes before they recombine for the output can be manipulated to perform matrix multiplication by creating a phase shift in the electromagnetic waves that make up light. Lightmatter has developed an energy-efficient, mechanical system of phase shifters called a Nano Optical Electro Mechanical System that uses electric fields applied to each arm of the MZI, which are made of a material whose refractive index will be determined by the electric field.[4] Creating the field only requires enough current to reach the intended field strength, after which point, no more energy is required.[4] The Mars chip, while highly energy-efficient and ideal for high-speed calculations, is currently limited to inference calculations. This means that the neural network must have already been trained in order for it to run the calculations.[4]

At the 2022 European Conference on Integrated Optics, George Washington University researcher Vorkel Sorger discussed how photonics offers the benefits of speed and low power consumption, with the goal of maximizing the operations per joule (OP J^{-1}). [9] Multiplication and addition operations performed in an electro-optical device are completed in picoseconds (10^{-12} s), compared to the already impressive nanosecond (10^{-9} s) speed that electronics usually require. Sorger and his colleagues have developed a photonic tensor core that can process signals at rates up to 100 TOP J^{-1}, which is faster than a typical commercial GPU.[9] This level of performance can significantly improve the applications of the tensor core to LiDAR and augmented reality, among other uses. Working with light, however, means there is the risk of optical loss, which must be reduced as much as possible to optimize the computing process.[9] While this technology significantly reduces costs coming from energy use and components, the majority of expenses come from the costs of testing and packaging.[9]

In theory, the applications of optical neural networks mimic the artificial neural networks they aim to replace. These applications include medicine, language processing, gaming, and image analysis, to name just a few.[13] Image classification has already been targeted as a promising application for existing optical networks. As Ashtiana, Geers and Aflatourni report, traditional processors for these networks are clock-based. Examples include GPUs, as mentioned above, and application-specific integrated circuits (ASICs).[10] GPUs can only compute as fast as their clock frequency allows, and ASICs, while more efficient in performance, still have shortfalls, including the fact that the dataset memory storage unit compromises information security. They recognize the high speed and efficiency of photonics, but also point out the lack of progress in developing scalable, on-chip, fully integrated photonic deep learning neural networks.[10] Their solution has the network's neurons all have the same range of output, meaning several layers of neurons can be implemented within the same chip. When tested for image classification accuracy, the chip's system was able to classify characters in handwritten letters with approximately 90% accuracy.[10] The process took only 570 ps, which rivals the speed of some of the best digital networks. This chip eliminates many of the components normally required for image detection and data processing, further improving efficiency, and the entire network can fit onto a 9.3 mm^2 chip.[10] Ashtiana et. al's chip faces a similar restriction to Lightmatter's Mars chip, in that the neural network has to first be trained at least partially off-chip to find the weight vectors of the neurons. Thus, a digital network must still be employed in order for the image classification to be possible.[10]

This caveat is common for optical neural networks. The reliance on a secondary neural network that can train the chips limits the feasibility of photonic chips to be standalone hardware.[12] For chips dedicated for a particular task (i.e. image classification), this is not a significant issue. However, if the photonic networks can not be adapted to perform changing tasks, they will be unable to fully replace traditional neural networks, an issue which starts to take away from the benefits of using photonics as an efficient alternative technology. Most neural networks are trained by a gradient-based backpropagation (BP) algorithm, but this same algorithm cannot be applied to optical neural networks.[12] Some experimental solutions to this issue have been proposed, including global optimization algorithms and forward propagation (FP) algorithms,[12] although these approaches are still only experimental at best.

Looking ahead:

With the exciting recent advancements in photonic chip technology, it is easy to view optics as the saving grace of computing hardware. Though it may be some time before electronics are fully replaced, it is worth considering the long-term consequences of reliance on any particular technology. In the not-so distant past, transistor-based computer chips were seen as groundbreaking contributions to computing. Moore's Law, though appropriate for several decades after its conception, could not be sustained indefinitely, as is becoming clear now. As photonics continues to gain traction as an efficient alternative to electronics, one must be cautious about pushing photonic technology to its limit at too rapid a pace. The high demands placed on neural networks' capability to tackle more intense and complex processes led to the need for alternative hardware in the first place. It is pertinent to use the current situation as an example for engineers, manufacturers and commercial buyers to consider the strain their expectations will put on resource availability.

In summary, photonics and optical neural networks offer a promising analog alternative to traditional artificial neural networks that rely on electrons to perform calculations for deep machine learning. Over time, it has become apparent that the energy and computing resources required to increase the performance of deep learning networks is no longer sustainable. Photonic chips use light to perform linear algebra calculations, a process that is both rapid and energy-efficient, and can execute tasks like image classification with high accuracy and performance speed. While these chips have already gone into production, like Lightmatter's Mars chip, a major roadblock in the path to switching to optical neural networks entirely is the stipulation that many of these prototype chips cannot be directly trained, but rather, the networks are trained off-chip, and the desired computing task is done by the chip's own network afterwards. Now that there is proof-of-concept for these chips' ability to perform deep learning calculations with the desired speed and efficiency, the next step for researchers and manufacturers will be to determine a method for on-chip training that does not sacrifice these advancements.

Literature

1) Hamerly, Ryan. “The Future of Deep Learning is Photonic.” https://spectrum.ieee.org/the-future-of-deep-learning-is-photonic. Accessed 8 September 2022.

2) Choi, Charles Q. “Photonic Chip Performs Image Recognition at the Speed of Light.” https://spectrum.ieee.org/photonic-neural-network. Accessed 10 September 2022.

3) Schneider, David. “Deep Learning at the Speed of Light”. https://spectrum.ieee.org/deep-learning-at-the-speed-of-light. Accessed 10 September 2022.

4) Schneider, David. “Lightmatter’s Mars Chip Performs Neural-Network Calculations at the Speed of Light.” https://spectrum.ieee.org/lightmatter-mars-photonic-chip-neural-network-calculations-speed-of-light. Accessed 9 September 2022.

5) Chen, Sophia. “Photonic Chips for Neuromorphic Computing.” https://spie.org/news/photonic-chips-for-neuromorphic-computing. Accessed 10 September 2022.

6) Cong, G., Yamamoto, N., Inoue, T. *et al.* On-chip bacterial foraging training in silicon photonic circuits for projection-enabled nonlinear classification. *Nat Commun* 13, 3261 (2022). https://doi.org/10.1038/s41467-022-30906-3

7) Vandoorne, K., Mechet, P., Van Vaerenbergh, T. *et al.* Experimental demonstration of reservoir computing on a silicon photonics chip. *Nat Commun* 5, 3541 (2014). https://doi.org/10.1038/ncomms4541

8) Zhu, H.H., Zou, J., Zhang, H. *et al.* Space-efficient optical computing with an integrated chip diffractive neural network. *Nat Commun* 13, 1044 (2022). https://doi.org/10.1038/s41467-022-28702-0

9) Pitruzzello, G. New frontiers for integrated photonics. *Nat. Photon.* 16, 559–560 (2022). https://doi.org/10.1038/s41566-022-01049-0

10) Ashtiani, F., Geers, A.J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* 606, 501–506 (2022).

https://doi.org/10.1038/s41586-022-04714-0

11) Xu, X., Ren, G., Feleppa, T. *et al.* Self-calibrating programmable photonic integrated circuits. *Nat. Photon.* 16, 595–602 (2022). https://doi.org/10.1038/s41566-022-01020-z

12) Britannica, The Editors of Encyclopaedia. "Moore’s law". *Encyclopedia Britannica*, 2 Sep. 2022, https://www.britannica.com/technology/Moores-law. Accessed 11 September 2022.

13) Hamerly, R., Bernstein, L., Sludds, A., Soljacic, M., Englund, D. Large-Scale Optical Neural Networks Based on Photoelectric Multiplication. *Physical Review X* 9, 21-32 (2022). https://journals.aps.org/prx/abstract/10.1103/PhysRevX.9.021032’