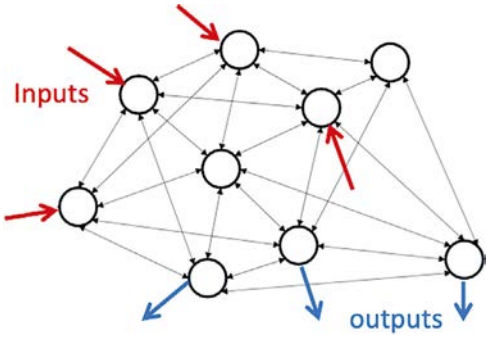
ARTIFICIAL INTELLIGENCE: FROM ELECTRONICS TO OPTICS

Sylvain GIGAN[1,3,*], Florent KRZAKALA[2,3], Laurent DAUDET[3], Igor CARRON[3]

[1] Laboratoire Kastler Brossel, ENS-Université PSL, CNRS, Sorbonne Université, Collège de France, Paris, France
[2] Laboratoire de Physique de l'École Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, Paris, France
[3] LightOn, 2 rue de la Bourse, Paris, France
***sylvain.gigan@lkb.ens.fr**



Machine Learning and big data are currently revolutionizing our way of life, in particular with the recent emergence of deep learning. Powered by CPU and GPU, they are currently hardware limited and extremely energy intensive. Photonics, either integrated or in free space, offers a very promising alternative for realizing optically machine learning tasks at high speed and low consumption. We here review the history and current state of the art of optical computing and optical machine learning.

https://doi.org/10.1051/photon/202010449

Since the dawn of micro-electronics and the emergence of lasers, both optics and electronics platforms have been competing for information processing and transmission. While electronics has been overwhelmingly dominating computing for the last 50 years thanks to Moore's law, optics and photonics have been increasingly dominant for communications, from long distance communications with optical fibers to optical interconnects in data centers. Machine Learning, that also originated in the 1950s, has seen tremendous developments in the last decade. The emergence of deep neural networks has become the *de facto* standard for big data analysis and many of the tasks that we today consider normal: from voice recognition to translation, image analysis to future self-driving cars. However, machine learning's progress requires exponentially increasing resources, be it in memory, raw computing power, or energy consumption. We introduce in this article the basics of neural networks, and see how this new architecture shatters the *status quo* and provides optics a new opportunity to shine in computing, whether in free space or in integrated photonics circuits.

EXPECTED CONTENTS

Optical computing. Classical computing, such as the one running on our PC, is based on the so-called Von-Neuman architecture laid out in the 1940s, where a program is stored in a memory, and instructions are read and executed on a processor, while input and output are exchanged in the memory through a communication bus. This architecture has been basically unchanged since its inception, and only improved thanks to the progress of microelectronics and nanolithography, allowing the feature sizes of components to shrink to 7 or less nanometers nowadays. This has consequently diminished tremendously the Ohmic losses and the energy consumption to a few pJ/operation, and allowed the increase of the operating clock frequencies of the components to reach several GHz. Thus, component density has driven the number of transistors on a processor to several tens of billions, while driving its cost down. This is the well-known Moore's law, leading to the observation that a good desktop PC nowadays has a processing power of several TeraFlops (10^{12} floating point operations per second).

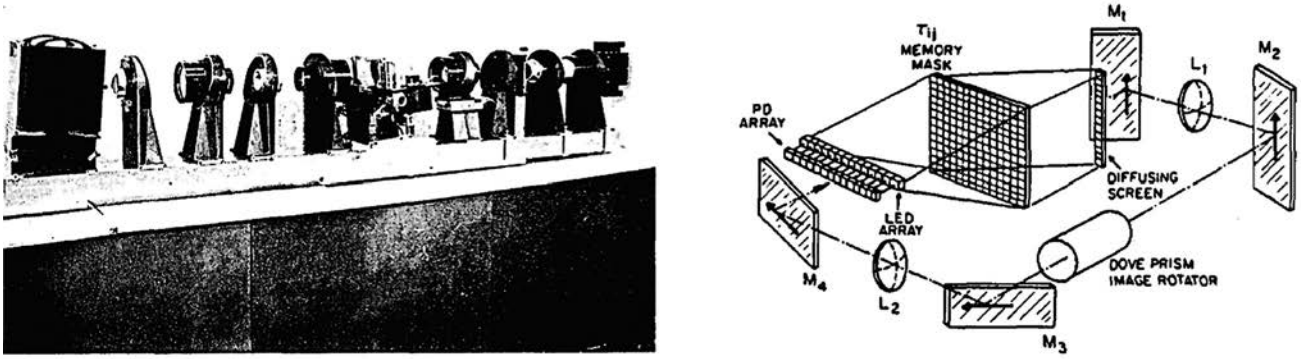


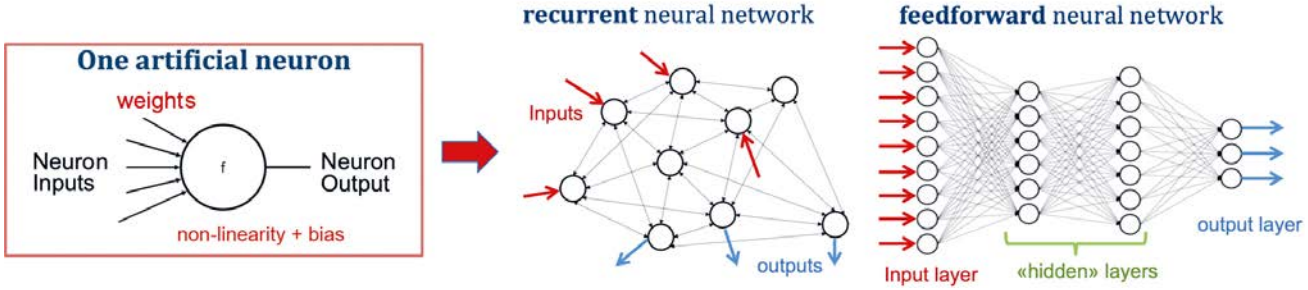
Figure 1.
Some historical examples of optical computing. Left: the 1972 Tilted Plane Optical processor used for synthetic Aperture Radar all optical image processing (from Kosma *et al.* Applied Optics 11.8 (1972): 1766-1777), right, a vector-matrix-multiplier with optical feedback (from Psaltis *et al.* Optics Letters **10.2** (1985): 98-100) Reprinted with permission from © The Optical Society.

Optics has several advantages compared to electronics: its intrinsic parallelism, its almost unlimited bandwidth, the ability of simple transformation by simple propagation (such as a Fourier Transform) compared to electrons [1]. Thus, optics has from the very start been considered as a viable alternative for analog computing. In the eighties, the emergence of optical non-linearities, semi-conductor lasers and optical memories has given hope that optics may be used to build an all-purpose computing platform. Alas, the progress in optics failed to match Moore's law exponential pace, and the hope of building such an all-purpose optical computer was abandoned in the nineties [2]. Still, optics found numerous applications in storage, and of course in telecommunications, both in long-distance with optical fibers, and more recently in interconnects.

Neural networks. In parallel, a computing paradigm, resolutely different from conventional programming, emerged also in the 1950's: Artificial neural networks or ANN, on which all modern artificial intelligence is based. It is (loosely) inspired from the structure and behaviour of the brain, where neurons are connected to each other in very complex networks, and where the response of a neuron can be triggered in a complex and non-linear way by the electric influx it receives from many other neurons. Artificial neural networks are similarly made of "neurons" or nodes, that integrate signals from other neurons, with various weights, and emit a resulting signal based on a non-linear activation function. This signal is, in turns, fed to a number of other neurons.

Figure 2.
Structure of an artificial neural network. Left: an artificial "neuron" comprising several inputs value, and one or many outputs, result of the non-linear combinations of the inputs. Center and left: two popular ANN architectures.

The network also includes input and output neurons, that either receive or send information to and from the outside world. Just like the brain, a neural network can be made to "learn", *i.e.* be optimized for a given task, by adjusting its weights, for instance being fed at the input with images, being able to classify them into categories. The analogy with the brain stops there: while the brain counts approximately 80 billions neurons and 100 trillions connections, ANNs have to be limited to much less neurons and weights, and to much simpler architectures, in order to make the training of the network possible. Several typical architectures have been developed over the last decades, to maximize efficiency on a given task, while keeping the training of the neural network computationally tractable. Most of the time, neurons are organized in layers of various sizes (number of neurons) and connectivity. It ranges from the simplest networks, such as the perceptron (a single layer linking N inputs to a single output) which was one of the earliest ANN, to multi-layered feedforward neural networks (where neurons are organized in successive layers and information is passed from



layer to layer) to recurrent networks (where information can flow backwards and be fed back to previous layers). The connections from a layer to the next can be very sparse, in particular convolutional layers, or dense (all-to-all connected). The performance of ANNs depends on its structure, for instance a perceptron is good for simple linear classification, but more complex tasks require more complex network structures.

Deep learning. While artificial intelligence saw good progresses, until the early 2000s, its overall performance for day-to-day tasks remained modest and did not find any clear real-life applications. This changed tremendously in the last two decades, thanks to the emergence of a powerful architecture: Deep learning, and its corollary networks, known as Deep Neural networks. Deep Neural networks are layered networks with a large number of "hidden" layers. Pioneers such as Yann Lecun, Yoshua Bengio and Geoffrey Hinton, have shown that deep neural networks have an ability to solve highly complex problems [3]: in essence, while the first layers can pre-process the input information (for instance contours in images, or words in text), deeper layers can gradually distil more abstract concepts, such as identifying an object, or extracting the sense of a text. Nowadays, deep learning has demonstrated unprecedented performance at tasks that we only recently believed would be forever out of reach of machines, from beating the best player at the game of Go, to self-driving cars, to language translation, to give just a few salient examples. Such deep networks have grown to unbelievable sizes, up to tens of billions of parameters (weights) to be trained. Thus, a key-enabling concept that has allowed deep learning to scale to large size is the ability to train such large network efficiently: the back-propagation algorithm, a concept perfectly matched to deep architectures, where the network can be trained layer by layer from the last to the first with a gradient-descent algorithm. Thanks to these, machine learning has entered the ability to make sense of complex and very large size information; this is sometimes coined as "big data".

GPU and CPUs. The rise of deep learning and big data has been mostly powered by Moore's law, allowing training and inference of very large neural networks. An important factor driving deep learning is the transition to Graphic Processing Units (GPU). Initially designed for computer graphics, these specialized processors were optimized for parallel processing of large vectors and matrices. For neural networks, where training and inference require a vast number of such multiplications, GPUs turned out to be much more powerful than CPUs (Central Processor Unit) and are now ubiquitous - incidentally, NVIDIA, the leader in GPU for deep learning, has now a capitalization that is on par with Intel. However, GPU and CPU are still enormously power-hungry: it has been shown that training a single neural network can use as much energy as 5 cars over their lifetime, and more globally, big data and data centers already account for an estimated 4% of our energy, and it may grow to over half of our energy consumption in the next decade, if nothing changes. Meanwhile, Moore's law is officially stalling: nanolithography and transistors are reaching their physical limits, progresses in consumption and speed are getting much slower [4]. Worse, the implementation of neural networks on both CPU and GPU suffers from the so-called "Von Neumann bottleneck": the bus transferring data between memory and computing units ultimately limits performance.

The dawn of optical Machine Learning. To overcome this fundamental problem, some non-conventional computing hardware has been introduced, called "neuromorphic", where circuits are directly emulating the connectivity and functions of a neural network, instead of a program on a CPU or GPU. This approach, that broadly belongs to non-von-Neumann architectures, should be much more energy efficient, and fast. Of all the possible implementations of neuromorphic computing, Optics and Photonics stands out, with unique advantages. First, light can propagate virtually without loss or heating, whether in free space, in many materials, or in integrated waveguides. This propagation

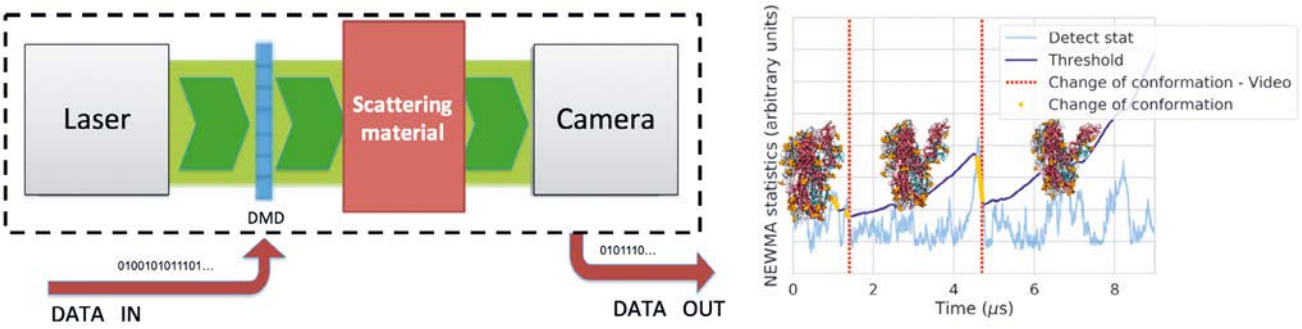


Figure 3.
LightOn's optical processor. Left: scheme of the random projections principle, information is encoded on a spatial light modulator, then a random matrix multiplication is achieved by passing through a disordered material, and the result is read of a camera sensor. Right: example of an advanced machine learning task, here the automatic detection of conformational change of a large molecule in molecular dynamic calculations (here on SARS-Cov2 molecule, responsible for the COVID-19 disease) [5].

can be used to emulate the connectivity between two neural layers, but also convolutions, etc. Photons do not naturally interact, meaning it is possible to multiplex information, and power consumption is independent of the operating frequency. Finally, thanks to tremendous progress in optoelectronics, detectors (from fast photodiodes to CMOS cameras), modulators (from fast integrated electro-optics modulators to spatial light modulators), and source (lasers), are extremely efficient and can be mass-produced. The semiconductor industry naturally provides the backbone to produce photonic integrated circuits. In short, optics has several key-advantages to implement neural networks in a nearly ideal way. Still, optics faces several challenges, in particular the difficulty to achieve non-linearities in hidden layers, or the challenge to scale and tune networks with integrated optics, preventing the possibility, to date, to provide a true versatile platform for deep learning. Yet, optics can provide a very solid alternative in specialized implementations, from ultrafast small scale networks, to convolutions and pre-processing in imaging, to reservoir computing (a type of RNN with fixed weights). After pioneering works in the 80s and 90s, many impressive advances have been reported in academia in the last decade, and industry also shown a renewed interest, whether within big companies or through start-up creations.

An example, LightOn. As an illustration of how optics can benefit machine learning, LightOn (the company we co-founded in 2016) proposed a solution to perform optical machine learning, based on our experience in free-space light propagation in complex media. In essence, we currently provide very large-scale random matrix multiplication (corresponding to a dense all-to-all connectivity) between millions of inputs (spatial light modulator pixels) and millions of outputs (camera pixels). Able to operate at several kHz, it corresponds to doing several Peta-Operations per second (typical of supercomputers), with a matrix size that could not even be stored in the memory of a conventional computer, and with a consumption of a few tens of Watts. While apparently very specific, the operation we propose can be useful in many data processing applications, from inference to training [5], or even molecular dynamics (see Fig. 3). In fact, these random multiplications can be seen as universal compression engines, with performance guarantees that are well matched to the very statistical nature of modern machine learning. Of course, this is just one approach to optical machine learning, and other approaches, either based on free space or integrated optics, fixed or tunable weights, linear or non-linear effects, shallow or deep, also proposes various solutions to accelerate machine learning and support its future growth.

CONCLUSION

In conclusion, we have presented an historic perspective of optical computing and shown that, after having failed at proposing an all-purpose computing platform in the 20th century, optics and photonics have more recently emerged as very appealing solutions for hybrid hardware implementation of neural networks, able to sustain the growth in computing power and supersede electronics, beyond Moore's law. Optical neural networks have recently rebooted the interest in optical computing, and we believe it is just the beginning.

RÉFÉRENCES

[1] J.W. Goodman, Opt. Photonics News **2**, 11 (1991)

[2] R. Athale, P. Demetri Psaltis, Opt. Photonics News **27**, 32 (2016)

[3] Y. LeCun, Y. Bengio, G. Hinton, Nature **521**, 436 (2015)

[4] M.M. Waldrop, Nature **530**, 144 (2016)

[5] LightOn white paper "Photonic Computing for Massively Parallel AI". https://lighton.ai/wp-content/uploads/2020/05/LightOn-White-Paper-v1.0-S.pdf