Computing with light: photonic processors for artificial neural networks

By Dr Frank Brückerhoff-Plückelmann

The influence of artificial neural networks on our society, working methods and science is constantly growing. Concrete examples include language models such as ChatGPT from the US company OpenAI, which are successfully used to generate text and images and can even handle complex tasks that require logical reasoning. At the same time, deep neural networks such as Google DeepMind's AlphaFold, which can predict the complex structure of proteins, are opening up completely new possibilities in science and research. Both the fundamental work on neural networks and their direct use in science were honoured with the Nobel Prizes in Physics and Chemistry in 2024. While economic and ethical issues surrounding the integration of artificial intelligence (AI) into everyday life are often at the centre of public debate, the sustainable provision of the necessary computing resources is also a key challenge [1]. An impressive example is the training of a GPT-3 model with 175 billion parameters: it took almost 15 days on 10,000 Nvidia V100 graphics cards and consumed 1,287 MWh [2]- equivalent to the total power output of a large nuclear power plant over one hour. There is also a trend towards ever larger models, as the number of trainable parameters, especially in the area of language models, usually has a positive effect on performance. These enormous resource requirements make it clear that more efficient and powerful hardware accelerators are needed to make the widespread use of AI sustainable and also to enable further development of the models.

In order to develop dedicated hardware for future AI applications, it is worth taking a look at the past. The first work on the perceptron, a fundamental structure in AI models, dates back to the 1940s. Despite initial enthusiasm and occasional breakthroughs, the lack of computing power soon led to disillusionment. This led to several AI winters in which research and investment in artificial neural networks were severely curtailed. The fundamental problem is easy to understand: Neural networks, inspired by their biological model, are based on a few types of operations, mainly additions and multiplications, which need to be performed frequently. In order to calculate a model efficiently, these operations must therefore be carried out in parallel on a massive scale. This is in stark contrast to the then prevailing Von Neumann architecture, which is designed for general calculations and is based on sequential programme execution. A decisive breakthrough came in the mid-2000s with the idea of using graphics cards - originally developed for parallel calculations in video games - for artificial neural networks. Today, specialised AI accelerators such as Google's Tensor Processing Units are specifically optimised for these requirements. In addition to parallel computing, their focus is particularly on minimising memory access in order to maximise efficiency. Interestingly, many of these hardware optimisations have similarities to the way the human brain works. For example, synapses in biological neural networks serve both as a "memory" for the connection strength between two neurons and as a "computer" that "calculates" the excitation of a neuron with the "stored" value. In contrast, the classic Von Neumann architecture has a strict separation between the computing and storage units. This development motivates research into new AI hardware that is even more inspired by the human brain.

In my doctoral thesis, I pursued this goal with the help of light-based analogue computers. The two fundamental differences to conventional hardware - analogue instead of digital computing and optical instead of electrical signal processing - offer new possibilities for the development of neuromorphic systems. Analogue computers encode a number directly by a physical quantity, such as the brightness of a light pulse, instead of using an abstract binary coding scheme. This allows computing operations to be carried out directly on a physical level in the memory, thus avoiding energy-intensive data transfer. An example of this is an object with variable transparency that simultaneously serves as a memory by storing a value in its transparency and functions as a computing unit. When a light pulse passes through the object, it is attenuated - the output brightness corresponds to the product of the input brightness and transparency. The use of light for calculations offers particular advantages that are also used in fibre optic networks, among other things. The high bandwidth in the terahertz range enables large amounts of data to be processed in parallel at high speed. At the same time, the propagation losses within the circuit are considerably lower than in electrical systems, which drastically reduces the amount of cooling required. In addition, light offers numerous physical degrees of freedom that can be utilised specifically for new computing approaches. The use of optical noise to capture the uncertainty of AI models is particularly interesting. In my research work, I have therefore developed a conventional and a probabilistic computing architecture.

Insight into AI: What actually needs to be calculated?

In order to develop special processors for AI applications, you first have to understand how neural networks work. This is an exciting field of research in itself, as the human brain has not yet been fully researched - it is therefore not possible to simply replicate it. In biology, neural networks consist of neurons that are connected to each other via synapses. When a neuron is activated, it sends electrochemical impulses, known as spikes, to other neurons. The synapses change the strength of the signal depending on how strongly two neurons are connected to each other. If a neuron receives sufficient signal strength within a certain period of time, it can in turn become active and send out its own impulses. Many modern deep artificial neural networks work in a similar way, but without this temporal dynamic. They are based on the *universal approximation theorem*, which states that a neural network can simulate any mathematical function if it has enough neurons and/or layers.

Figure 1a shows the structure of a single-layer neural network consisting of an input layer with four neurons and an output layer with one neuron. For example, the neurons in the input layer of such a network could record the brightness values of the individual pixels of an image. Mathematically, the values of the neurons in a particular layer can be visualised as a vector with N components. The next layer is calculated by matrix-vector multiplication, followed by a non-linear activation function. This means that if the first layer has N neurons and the second layer also has N neurons, N^2 multiplications and N non-linear operations must be performed. Therefore, matrix-vector multiplications are by far the most computationally intensive and energy-consuming operations in neural networks. And therefore the greatest gain lies in the development of specialised hardware that can perform precisely these calculations faster and more efficiently.

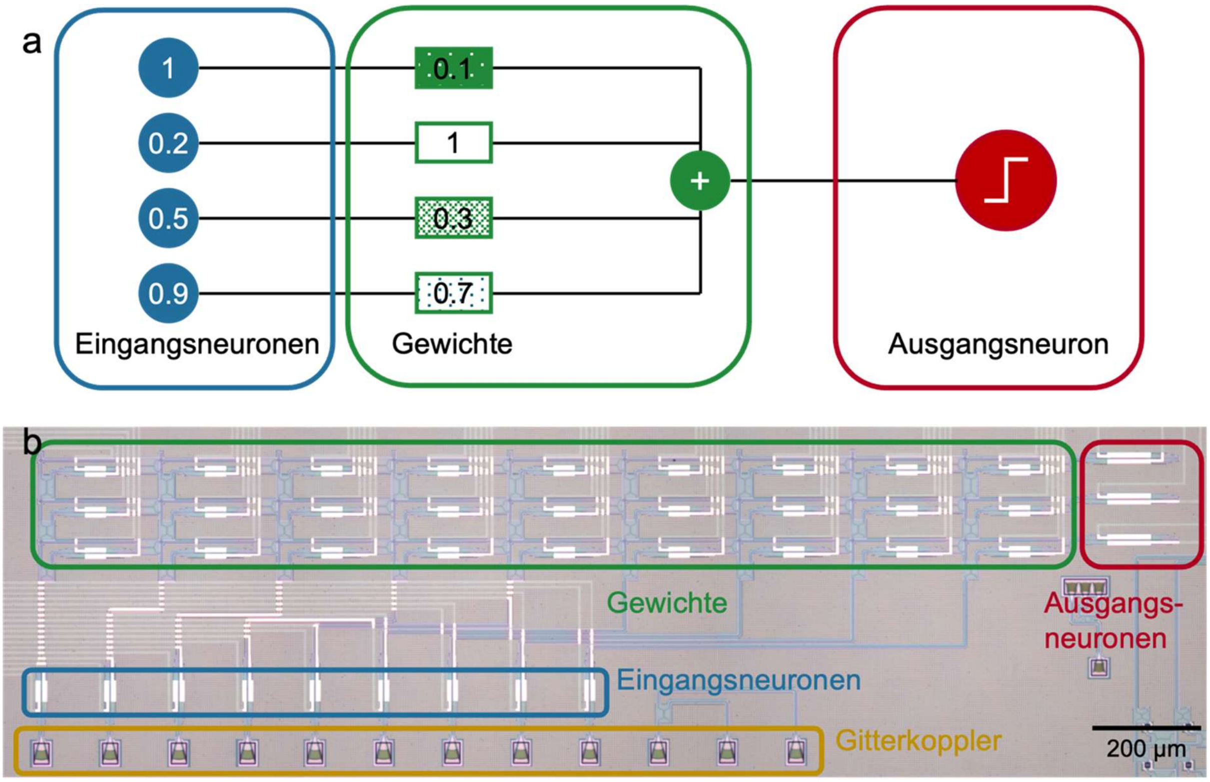

Figure1 . AI and photonic processors. ***a****, To determine the value of an output neuron, the activation of the input neurons is multiplied by the corresponding synaptic weights. The products are added and the activation of the output neuron is calculated by a non-linear function.* ***b****, The photonic circuit calculates the multiplications and additions for nine input neurons and three output neurons simultaneously. The grating couplers (yellow) couple the light into the chip, the input modulators (blue) encode the activation of the input neurons and the photodiodes measure the value of the weighted sum to calculate the output activation (red). The computational memory simultaneously stores the weights and calculates the multiplications and additions (green).*

In my research, I have developed a photonic circuit, see Figure 1b, that calculates matrix-vector multiplications with light [3], rather than with electrical pulses as in conventional processors. Similar to biological neural networks, information is represented by physical quantities - in this case by the brightness of light pulses and the transmission of adjustable absorbers. In addition, the computing operations are carried out directly in the memory, which offers a fundamental advantage over conventional systems. The photonic circuit I developed, which calculates a matrix-vector multiplication with nine input neurons and three output neurons, consists of several components. The grating couplers direct light from an external light source into the chip - comparable to a power cable that conducts electrical energy into a computer. Electro-absorption modulators then take over the control of the light intensity. These work at speeds of up to 50 GHz and encode the values of the input neurones directly into the brightness of the light pulses. A maximum brightness corresponds to a neuron value of 1, while a value of 0 means that the light is blocked to the maximum. The actual calculation process takes place in an interconnection of various photonic components, outlined in green in Figure 1b. This is where the input values are added and multiplied. The process works as follows: Each of the input light pulses is first split into three equally bright partial pulses. The brightness of each of these partial pulses is then attenuated individually by an electro-absorption modulator . The strength of this attenuation corresponds to the synaptic weights of the neuronal network. Finally, in each of the three lines, the individually attenuated light pulses are recombined into a single light signal. This signal is then read out using a photodiode to determine the result of the calculation. This photonic computing architecture differs fundamentally from conventional electronic systems. While conventional computers perform such multiplication and addition operations (sequentially) at clock rates in the gigahertz range, the speed in my circuit is limited only by the speed of light in the optical waveguides.

Demonstrator: Machine vision at the speed of light

The photonic circuit forms the centrepiece of the optical processor. However, it requires an interface to the predominantly digital and electronic environment in order to read data in and out. For electronic data exchange, the chip is bonded to an electrical circuit board. The chip's contact pads are connected to the circuit board via thin wires so that electrical signals can be transmitted (see Figure 2a). This circuit board connects the chip to digital-to-analogue and analogue-to-digital converters. This allows digital data to be converted into analogue signals, processed in the optical circuit and then transformed back into digital data. The optical coupling takes place via a series of optical fibres that are anchored in a glass block (illuminated in red in Figure 2a).

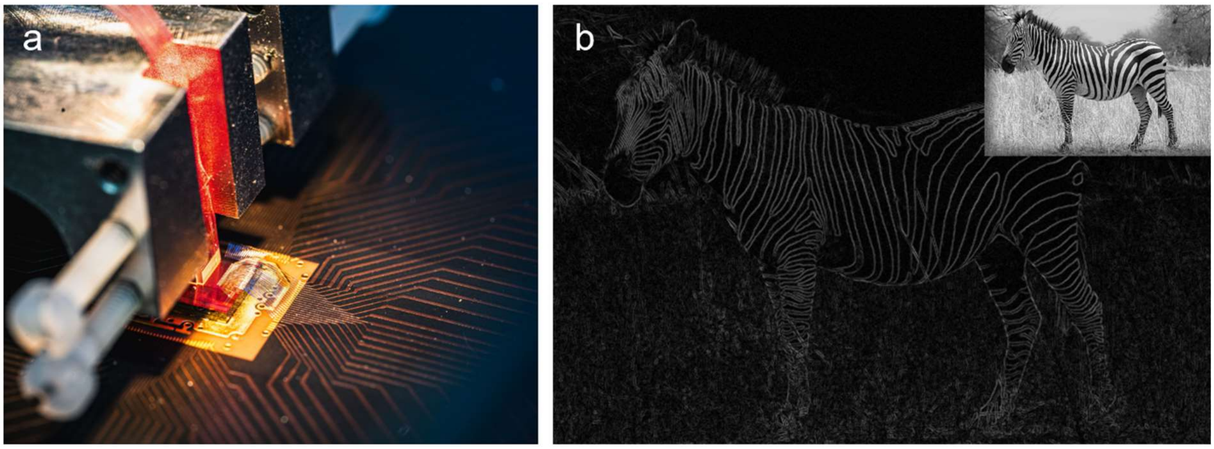

Figure2 . Edge detection with a photonic circuit. ***a****, The chip is bonded to a printed circuit board to control the photonic processor. In this way, electrical signals are sent to the chip and light is coupled into the circuit at the same time.* ***b****, Among other things, the chip can calculate folds for edge detection. The inset shows the input image.*

The optical processor can perform 2 billion matrix-vector multiplications per second [3]. Its maximum performance is currently limited by the electronic interface that provides the data transfer between the digital environment and the photonic circuit. As part of the EU research project Phoenics, among others, I used this demonstrator (experimental setup) to optically calculate so-called convolutions - matrix multiplications that are important for processing visual data. During convolution, a small numerical filter, also known as a kernel, is moved across the image to identify local features such as edges. This information is then further processed in the neural network, for example to classify objects in the image. In Figure 2b, an edge detection filter has been applied to the image of a zebra - all the necessary matrix-vector multiplications were calculated entirely with the photonic circuit. We even tested the system in a real-time scenario: A live video from a webcam was streamed directly into the optical processor to recognise the edges in the image in real time. This demonstrates the enormous potential of optical computing for future AI applications, especially in areas such as automated image processing and machine vision.

Utilising chaos: How optical noise makes AI more reliable and safer

In addition to applications within conventional neural networks, photonic processors also open up new computing operations. These make it possible to calculate other mathematical models efficiently - similar to how the use of graphics cards has revolutionised the development of artificial neural networks. One particularly interesting area is the more precise modelling of uncertainties. This would allow an AI to not only make predictions and give recommendations for action, but also provide an assessment of how certain it is in doing so. A typical example is behaviour in unknown situations. Humans intuitively know when they are confronted with a completely new, never-before-seen situation - and that they cannot make a reliable statement in such cases. Conventional AI models, on the other hand, can easily be misled and make false statements with a high degree of confidence, even with completely unknown inputs. This can have serious consequences, especially in safety-critical applications such as autonomous driving. Bayesian neural networks (BNNs) are one possible solution. Unlike in classic neural networks, in which the weights are stored as fixed values, in BNNs they are represented as probability distributions. This enables such a network not only to make a decision, but also to estimate how reliable it is. For example, it recognises whether it feels uncertain because it has never seen a certain input before or because the available data does not allow a clear prediction. In practice, however, BNNs require probabilistic computing, in particular the ability to draw random values according to probability distributions. Since classical processors work deterministically, they have to use large computing capacities with the help of pseudo-random number generators to generate the required random values - a computationally intensive and inefficient process. In my doctoral thesis, I developed a photonic probabilistic computer that generates these random values directly during the matrix-vector multiplications. It works with the same high optical data rates as the previously presented photonic chip for classical neural networks. The key to this lies in the spontaneous amplified emission within erbium-doped optical fibres, which inherently generate random power fluctuations. In combination with a specially developed coding scheme and a photonic circuit (similar to the one in Figure 1b), I was able to calculate probabilistic matrix-vector multiplications. As a practical test, we trained a photonic BNN to correctly classify handwritten digits from 0 to 8. In addition, it was confronted with the previously unseen number "9" - and successfully recognised it as an unknown input instead of making an incorrect assignment with high confidence [4]. This shows the potential of photonic computers for the next generation of secure and trustworthy AI systems.

Outlook: Photonic processors in every computer?

Photonic processors offer enormous potential for performing analogue calculations efficiently. They open up new possibilities, particularly in the field of artificial intelligence, as their architecture can be more closely modelled on the way biological neural networks work. This allows computing power to be increased and energy consumption to be reduced - two of the biggest challenges facing current AI hardware. In my PhD thesis, I developed prototypes of photonic processors that can be used for both conventional neural networks and probabilistic neural networks. Despite this progress, there are still challenges on the way to the widespread use of photonic computers. The biggest hurdle is the high development costs required to bring a new technology to market. Electronic architectures benefit from decades of investment, while photonic systems are only at the beginning of this development. For this reason, photonic processors will initially be used primarily in specialised high-performance applications where the computing effort is particularly high and conventional architectures have reached their limits. One promising example is the integration of photonic connections in data centres. Google is already using optical connections between computing chips to make data transfer more efficient on a large scale [5]. The integration of photonic components at chip level is being intensively researched, particularly in view of the increasing amounts of data that need to be processed in AI applications. Photonic computers could benefit from these advances, especially if more and more peripheral components are converted from electrical to optical interfaces. This would simplify the interface between conventional and photonic hardware, which would facilitate a wider range of applications. A particularly exciting future field is probabilistic computing, especially with Bayesian neural networks. These offer great theoretical advantages, but are not yet widespread as there is no truly efficient digital implementation. Physical analogue computers such as photonic systems are particularly interesting here, as they can use natural noise as a source of randomness instead of generating it artificially - a decisive advantage over classic, purely digital solutions.

Overall, my research shows that photonic computers have enormous potential, especially in combination with optical interconnects and the associated advances in scalable manufacturing methods. Development in this area is still in its infancy, but the course is set - and in the coming years, photonic processors could play a central role in the next generation of AI hardware.

References

1. Crawford, K. Generative AI's environmental costs are soaring - and mostly secret. **Nature** 626 (2024).
2. Patterson, D., Gonzalez, J., Le, Q. et al. Carbon Emissions and Large Neural Network Training. **ArXiv** (2021).
3. Dong, B.*, Brückerhoff-Plückelmann, F.*, Meyer, L. et al. Partial coherence enhances parallelised photonic computing. **Nature** 632 (2024).
4. Brückerhoff-Plückelmann, F., Borras, H., Klein B. et al. Probabilistic photonic computing with chaotic light. **Nat Commun** 15 (2024).
5. Jouppi, N. P., Kurian, G., Li, S. et al. TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. **Proc Int Symp Comput Archit** (2023).