Efficiency and scalability of optical neural networks

Michał Matuszewski and Andrzej Opala

Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46, PL-02-668 Warsaw, Poland

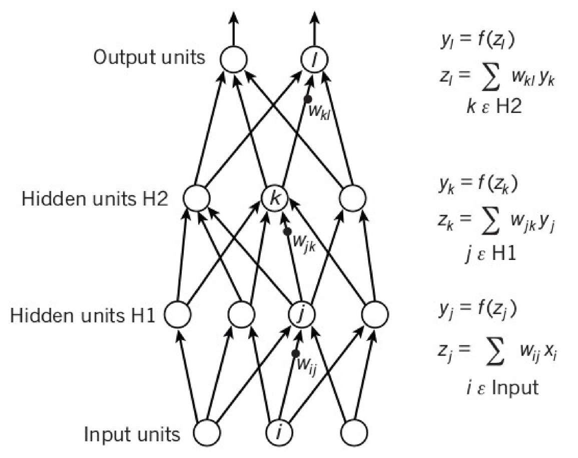
Remarkable developments in big data, artificial intelligence and neural networks come at the cost of high energy consumption that is necessary to process large amounts of data. In result, much research has been aimed at finding alternative platforms for information processing, characterized by high performance and energy efficiency. We consider the advantages of optical neural networks [1]. Neural networks is likely to be the first area where optical systems could outperform electronic specialized systems such as GPUs and TPUs. We discuss the advantages of exciton-polariton systems as the most promising candidates for all-optical nonlinear information processing [2-5].

Neural networks and the energy consumption problem

A large number of samples is used as an input for the teaching algorithm during the **training phase**

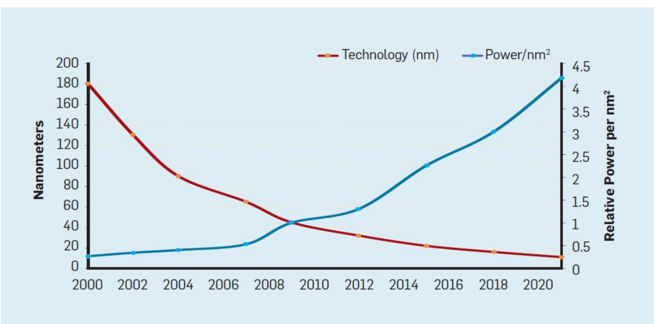
The algorithm produces a neural network with tuned weights, which is able to do the job in the **inference phase** not only on the training data, but also on samples that it has never seen before

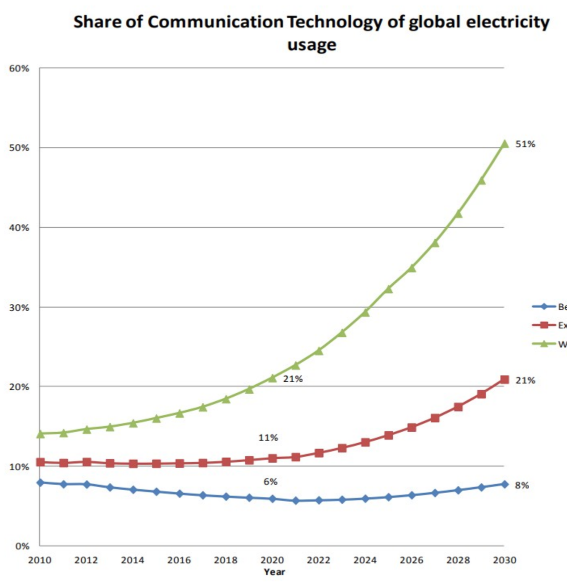
Y. LeCun, Y. Bengio, and G. Hinton, Nature 521, 436 (2015).



Energy consumption due to information processing and communications as a share of the global electricity usage is quickly growing

Electronics will not be able to meet the demands in the future as Moore's law and Dennard scaling come to an end



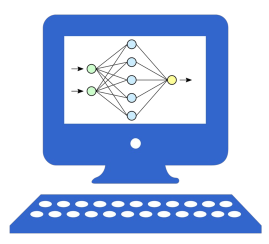


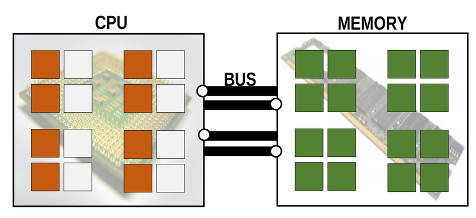
John L. Hennessy, David A. Patterson Communications of the ACM, February 2019, 62, 48-60
Anders S. G. Andrae and Tomas Edler, Challenges 2015, 6, 117-157

Neural networks contribute significantly to this problem

Current implementations of neural networks, based on computer simulations, suffer from the **von Neumann bottleneck** which is due to the limited capacity and efficiency of communication channels

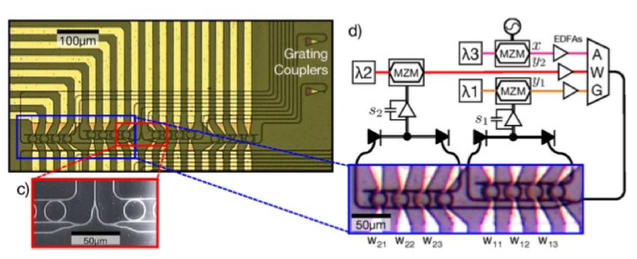
A possible solution is the development of **neuromorphic computing**, where the neural network structure is resembled in hardware





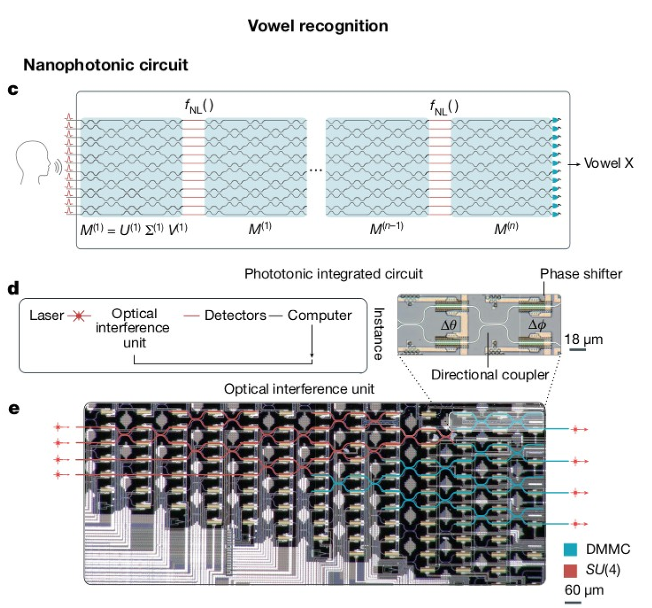
Optical neural network efficiency and scalability

Optical neural networks use a variety of designs that can be roughly divided into **optoelctronic** vs **all-optical** and **integrated** vs **free-space**



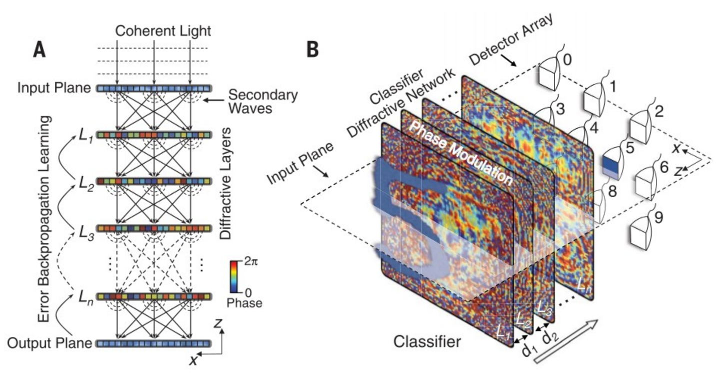
A. N. Tait et al., J. Lightwave Technol. 32 , 3427 (2014)

An integrated, optoelectronic network



X. Lin et al., Science 361, 1004 (2018)

A free-space network



Shen, Y. et al., Nat. Photon. 11, 441 (2017)

An integrated, all-optical network

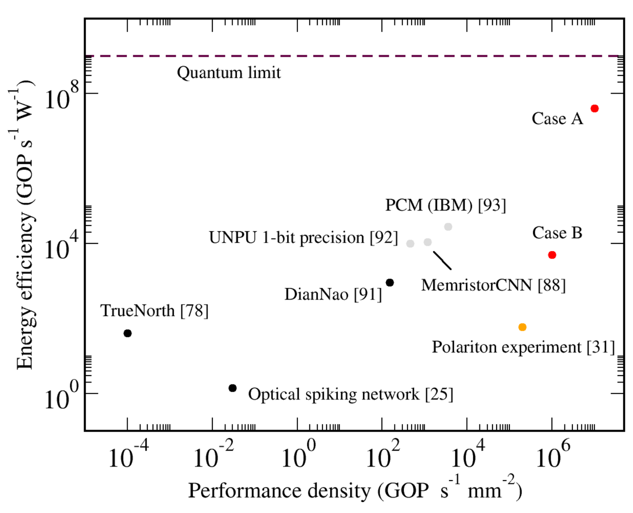
- Integrated networks can be more easily coupled to electronics, but incur waveguide losses
- Optoelectronic networks can realize arbitrary activation functions and provide optical signal regeneration, but are difficult to scale

We propose **free-space, all-optical polariton networks**.

To get a realistic estimate of efficiency, we take into account the energy cost of light source, modulators, detectors, and optical losses.

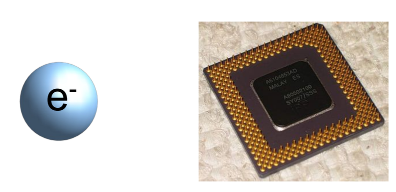
(A) an "idealized" large scale system, with parameters corresponding to state-of-the-art optical elements,
(B) a proof-of-principle system with a relatively small number of nodes and accessible optical elements.

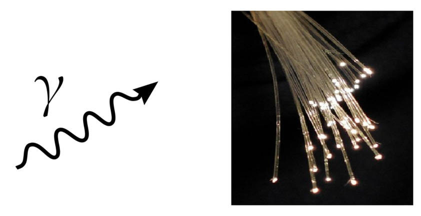
Orders of magnitude improvements over electronics in energy efficiency (aJ range) and processing speed per mm² are predicted



*M. Matuszewski et al., Phys. Rev. Appl. **16**, 024045 (2021)*

Polariton neural networks



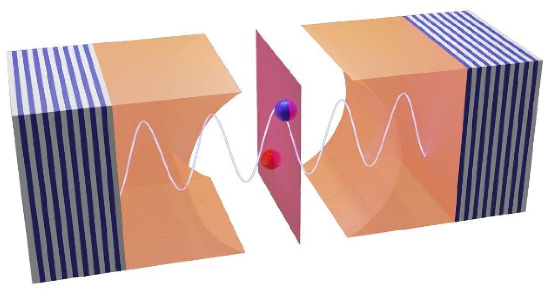


Electrons	Photons
Excellent for information processing	Excellent for communication of information
Substantial interactions	Interacting weakly, only in nonlinear media
High losses	Low losses

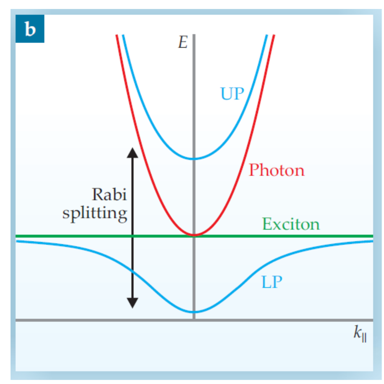
The weakness of interactions is a severe problem for optical computing

Exciton-polaritons

- Quantum eigenstates of strongly coupled excitons and cavity photons are quasiparticles called exciton-polaritons
- Excellent transport properties and extremely low effective mass thanks to the photonic component
- Extremely strong interparticle interactions thanks to the exciton component – world record of ultrafast optical nonlinearity

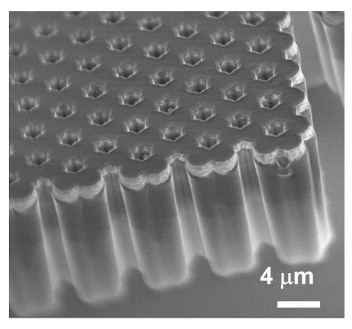


$$\hat{H}_k = E_{ex} a^+ a + E_{ph} b^+ b + \frac{\Omega}{2}(a^+ b + a b^+)$$



Proposal for polariton neural networks

*A. Opala, S. Ghosh, T. C. H. Liew, M. Matuszewski, **Phys. Rev. Applied 11**, 06402, (2019)*



Resonant injection — NN Coupling

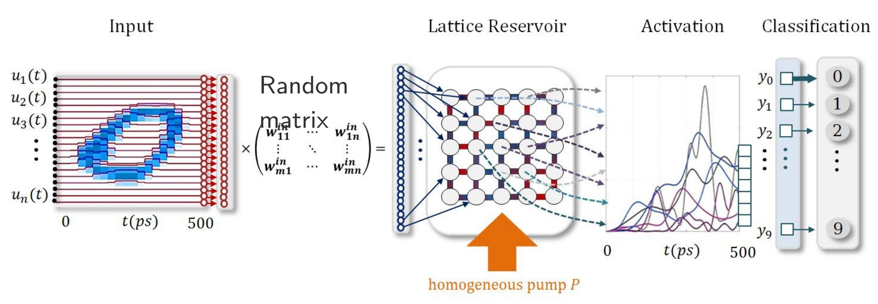
$$\frac{d\psi_n}{dt} = W^{in}_{nm} u_m - i \sum_{m=nn} W_{nm}\psi_m + \left(\gamma - \Gamma|\psi_n|^2 - ig|\psi_n|^2\right)\psi_n,$$

$\gamma = P - \kappa$

Nonlinear losses — Interactions

Polariton lattice is excited with a series of resonant pulses encoding input and a nonresonant background pump P.

We chose the reservoir computing architecture for the neural network, which includes a hidden layer of non-tunable, nonlinear polariton nodes with random connections

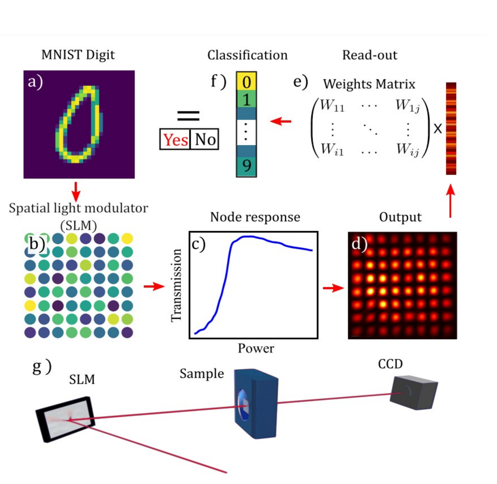


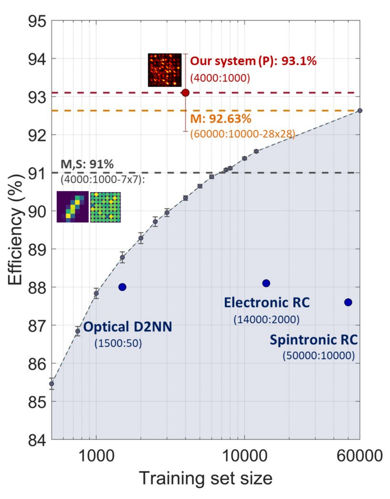
Reference	Technology	TI 46 word rate
Q. Wang et al., IEEE International Symposium on Circuits and Systems (ISCAS). pp. 361–364, 2016	CMOS FPGA	2 500 words / s
Bogdan Penkovsky, PhD thesis, Université Bourgogne Franche-Comté, 2017	CMOS FPGA	350 words / s
L. Larger et al., Physical Review X 7, 011015 (2017)	Electro-optic delay line	7.7 x 10^5 words / s
This work (numerical estimate)	Microcavity polaritons	1.6 x 10^{10} words / s

Experiments

The first experiment, realized at the CNR Nanotec in Lecce, followed our theoretical proposal, but data was encoded in space and not in time.

The achieved accuracy in the Modififed NIST handwriten digit dataset was 93%, significantly above linear classification benchmark and higher than in other neuromorphic systems.





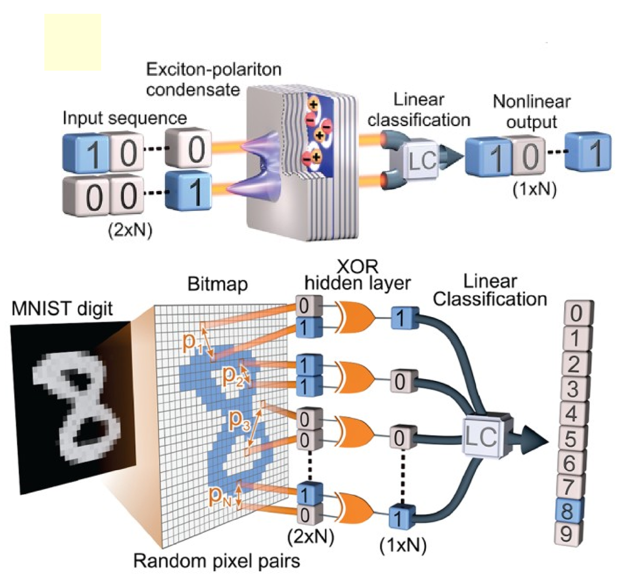
*D. Ballarini et al., **Nano Lett. 20**, 3506–3512 (2020)*

In collaboration with the University of Warsaw, we designed a binarized neural network where neurons are realized as polariton XOR gates

In the first, proof-of-principle experiment, linear classification was performed in software. **Nonlinear transformation** was realized entirely with optical elements

The achieved MNIST accuracy of 96% is similar as in state-of-the-art neuromorphic hardware realizations

The optical pulse energy per synaptic operation was 16 pJ, which can be compared compared to around 100 pJ per MAC in typical GPUs and 16 pJ in Frontier, the current #1 in the Green500 list.



*R. Mirek et al., **Nano Lett. 21**, 3715–3720 (2021)*

[1] B. J. Shastri et al., Nature Photonics **15**, 102 (2021).
[2] A. Opala, S. Ghosh, T. C. Liew, and M. Matuszewski, Phys. Rev. Appl. **11**, 064029 (2019).
[3] D. Ballarini et al., Nano Letters **20**, 3506 (2020).
[4] R. Mirek et al., Nano Letters **21**, 3715 (2021).
[5] M. Matuszewski et al., Phys. Rev. Appl. **16**, 024045 (2021).

This work was supported by National Science Center, Poland via Grant No. 2017/25/Z/ST3/03032 under the QuantERA program.