The 32nd Hot Chips Symposium

LIGHT-IN-THE-LOOP: USING A PHOTONICS CO-PROCESSOR FOR SCALABLE TRAINING OF NEURAL NETWORKS

Julien Launay [1,2], Iacopo Poli [1], Kilian Müller [1], Igor Carron [1], Laurent Daudet [1], Florent Krzakala [1,2], Sylvain Gigan [1,3]

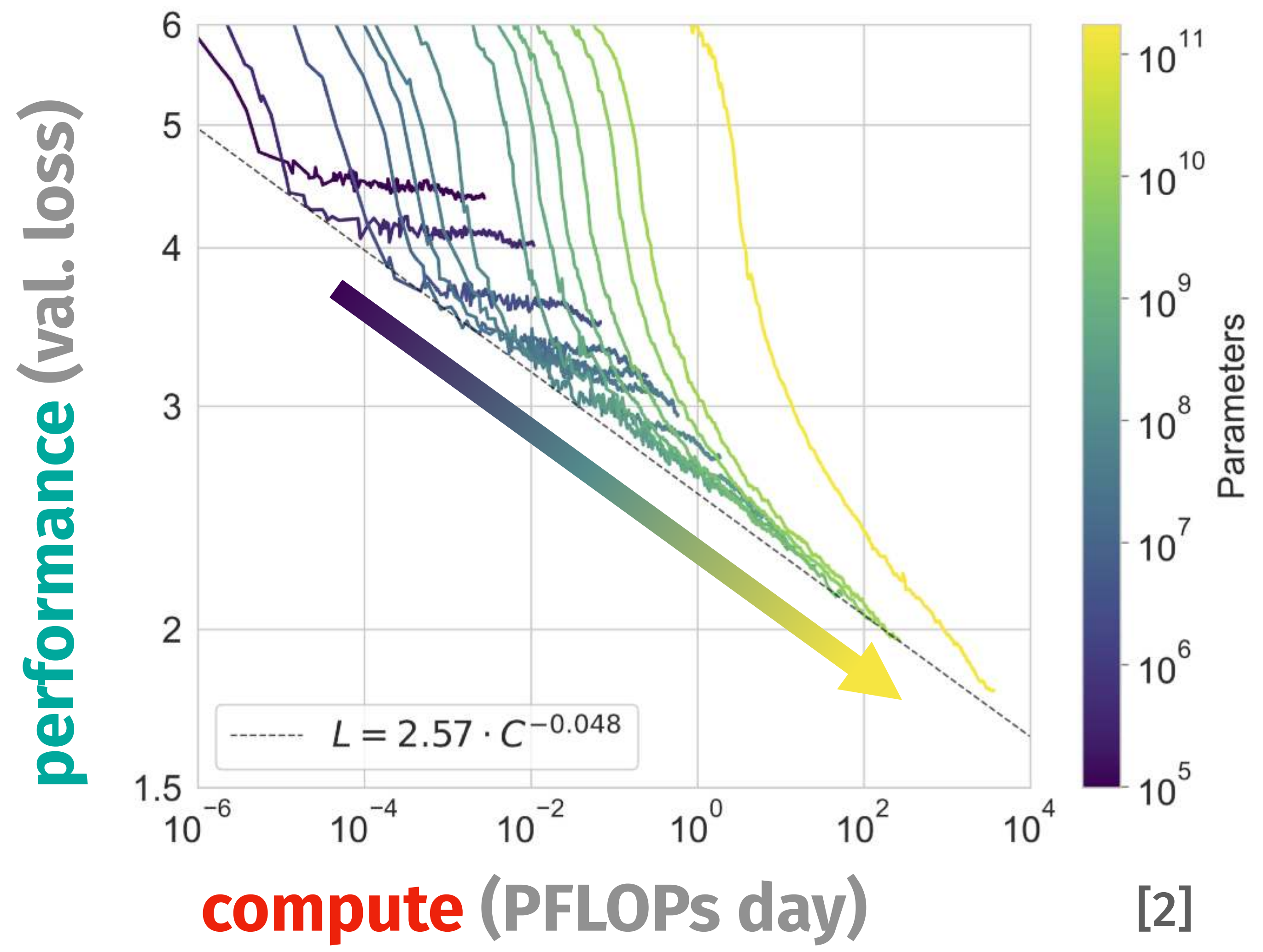
Contact: {firstname}@lighton.ai

[1] LightOn
lighton.ai

Laboratoire de Physique de l'École Normale Supérieure [2]

Laboratoire Kastler Brossel [3]

Scaling deep learning: more compute is all you need



- 🔥 Neural networks are growing larger... 175 billion trainable parameters in GPT-3![2]
- ✅ Staggering correlation between size and performance.[1]
- 👏 Enabling larger models will benefit deep learning.

🤔 How can we scale training hardware to such large models?

New co-design philosophy: hardware for learning beyond backpropagation

inform the design of...

novel training methods
alternatives to backpropagation
with compelling characteristics

dedicated hardware
advanced technologies targeting
specific operations and processes

...enables the use/scaling of

LightOn

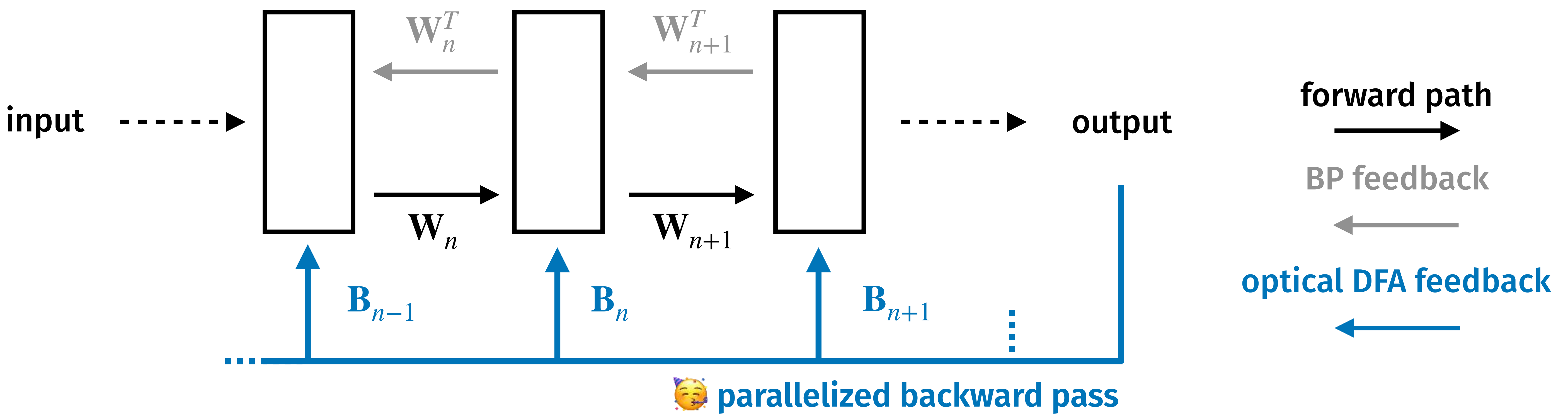
LightOn OPU: the first at scale photonic co-processor in the cloud

💪 On the hardware side...

- 💡 Multiple light scattering through a diffusive medium provides massively parallel random projections! [3]
- 🎓 Leveraging holography to retrieve a linear random projection.
- 📈 1Mx300k random projection in 1.25ms with less than 40W... with expected x10 performance every 2 years.
  up to 2 trillions parameters per RP, equivalent to ~1 petaOPS with 2TB cache memory in a non-von Neumann architecture.

🧠 On the training side...

- 💡 Direct Feedback Alignment (DFA)...[4] scales to modern deep learning tasks and architectures.[5]



Replace BP updates,

$$\delta \mathbf{W}_i = -[(\mathbf{W}_{i+1}^T \delta \mathbf{a}_{i+1}) \odot f_i'(\mathbf{a}_i)]\mathbf{h}_{\mathbf{i-1}}^{\mathbf{T}}$$

with a random projection of the error,

$$\delta \mathbf{W}_i = -[\underbrace{(\mathbf{B}_i^T \mathbf{e})}_{\text{performed with light}} \odot f_i'(\mathbf{a}_i)]\mathbf{h}_{\mathbf{i-1}}^{\mathbf{T}}.$$

🎉 Photonic training demonstrated on MNIST: more coming soon, holographic photonic core OPU pre-release end of 2020.

Light-in-the-loop: poster overview

slides

5-7 Training neural networks with Direct Feedback Alignment

8-9 Holography and photonic for linear random projections

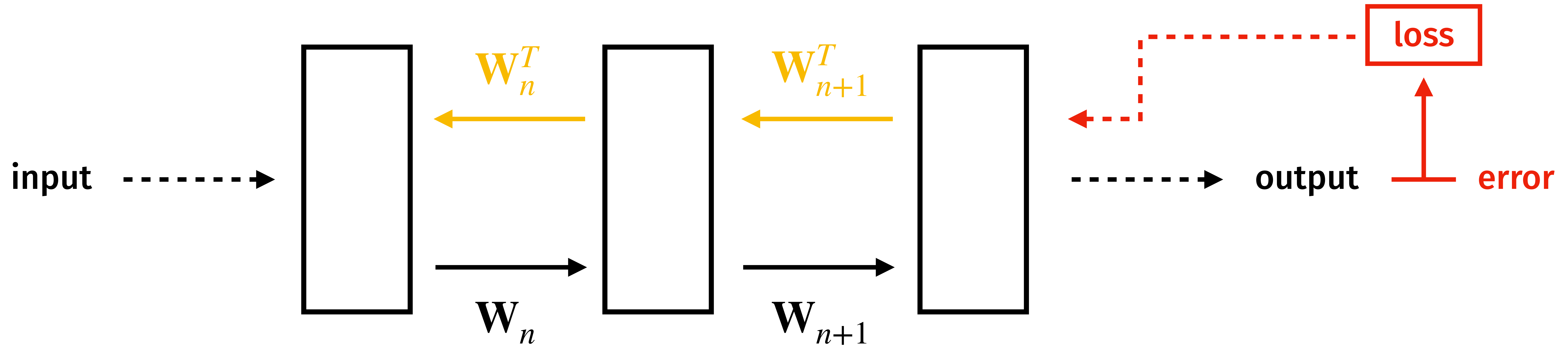
10 Light-in-the-loop: photonic training of neural networks

Training neural networks the usual way: backpropagation of the error (BP)

💳 Training a neural network is a credit assignment problem: find which neurons are to blame for the error at the output.

🚨 Backpropagation is the canonical training method, and assigns blame precisely to each neuron.

🙅 But doing so brings practical limitations (e.g. no parallelization of the backward pass) and is not biologically plausible.[6]



Can we train neural networks differently?

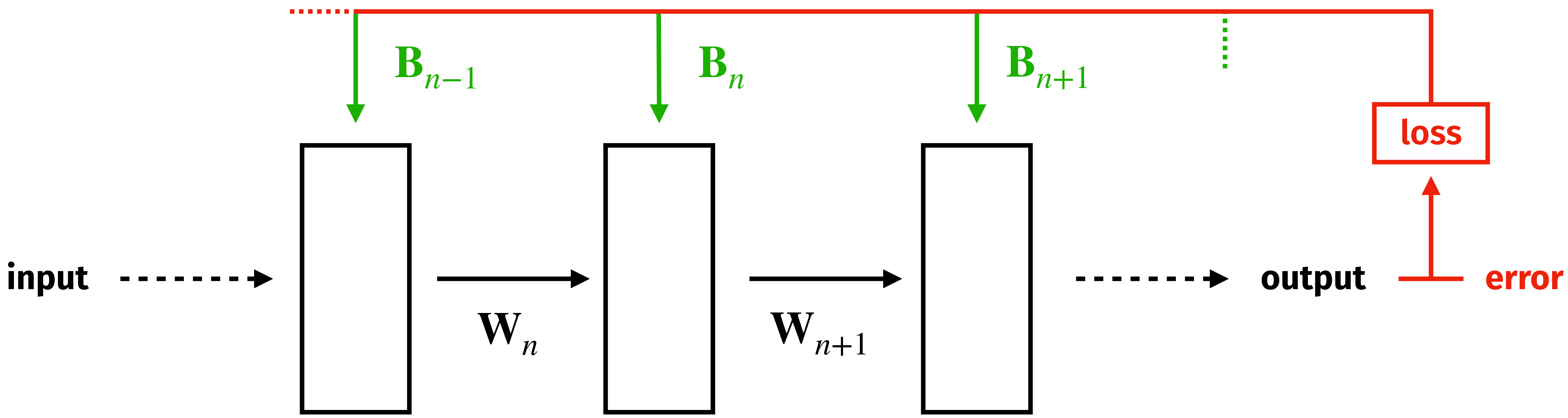
alternative methods can combine
computational and biological motivations

LightOn

Training neural networks with... direct feedback alignment (DFA)[4]

🧠 **Biologically-inspired (weight transport problem): uses random weights to deliver feedback from global loss.**

💪 **Puts a single random projection at the cornerstone of training and enables parallelization of the backward pass.**



gradient signal from layers above — random projection of the loss gradient

$$\delta \mathbf{W}_i^{BP} = -[\overline{(\mathbf{W}_{i+1}^T \delta \mathbf{a}_{i+1})} \odot f_i'(\mathbf{a}_i)]\mathbf{h}_{\mathbf{i-1}}^{\mathbf{T}} \longrightarrow \delta \mathbf{W}_i^{DFA} = -[\overline{(\mathbf{B}_i^T \mathbf{e})} \odot f_i'(\mathbf{a}_i)]\mathbf{h}_{\mathbf{i-1}}^{\mathbf{T}}$$

🎓 **Mathematical intuition:** $(\mathbf{B}_i^T \mathbf{e})^T (\mathbf{W}_{i+1}^T \delta \mathbf{a}_{i+1}) > 0 \Rightarrow$ **DFA update within 90° of BP**

(going roughly in the right direction is enough)

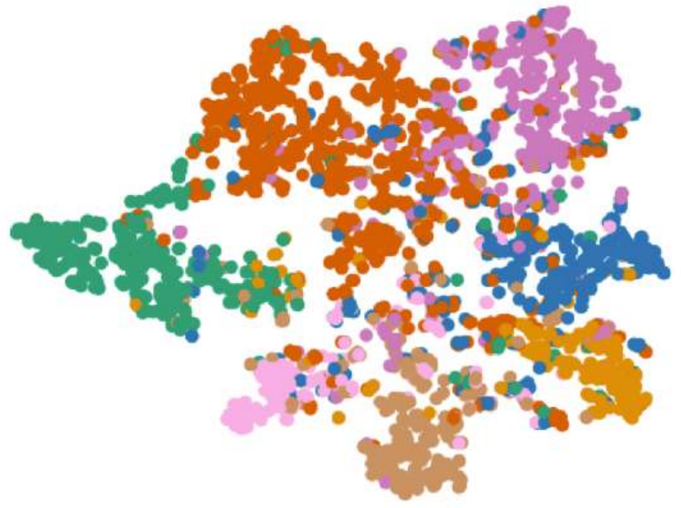
can be enforced by tuning $\mathbf{W}_{i+1}$: *learned* alignment of the forward weights.

LightOn

DFA scales to modern deep learning tasks and architectures

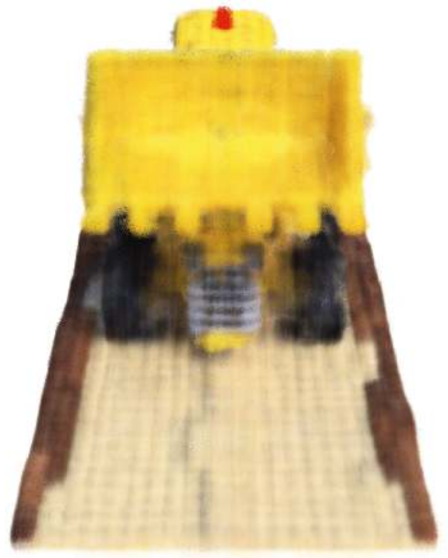
🙋 Is DFA as versatile as BP as a training method? [5]

✅ Graph Convolutional Neural Networks



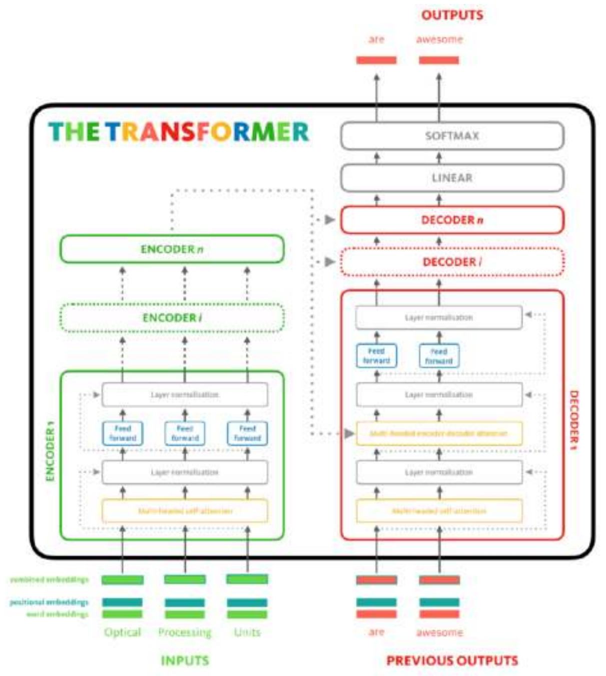
t-SNE of graph embeddings learned with DFA.

✅ Neural scene modeling with NeRF



Synthetic 3D scene learned with SOTA methods.

✅ Massive hybrid Transformer architectures



DFA and BP can be mixed to accelerate the training of large architectures with limited accuracy cost.

🧐 One caveat: DFA doesn't work on convolutional layers (yet!) [7, 8]

📝 Further research work is needed to improve theory of DFA and build principled training methods.

BP has its own tricks: dropout, batch normalization, etc.

👌 Hybrid strategies (DFA+BP) can work out of the box in large architectures.

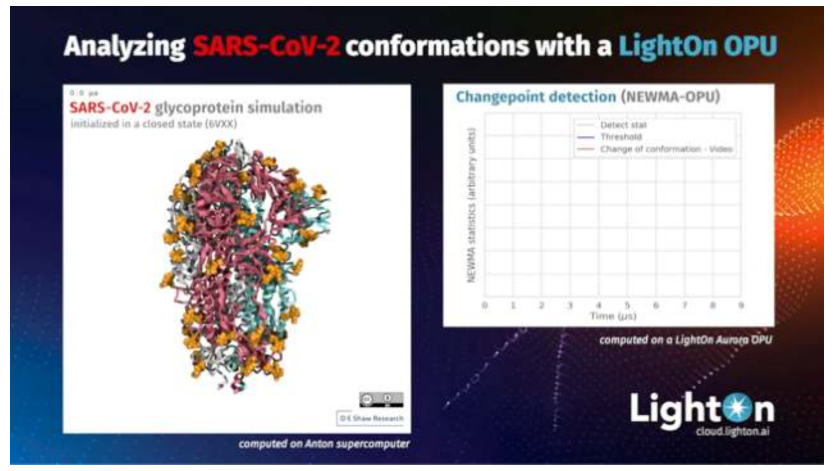
LightOn OPU: a photonic co-processor for random projections

💡 Leveraging photonics brings a number of advantages:

- 🔀 Massively parallel processing, with the entire projection computed at once; — equivalent to millions of cores
- ♾ Very high dimensional input and output, with easy scaling; — up to 1Mx2M random projection
- ⚡ Energy-efficient hardware, as the computation is mostly passive. — only 30W for the packaged system

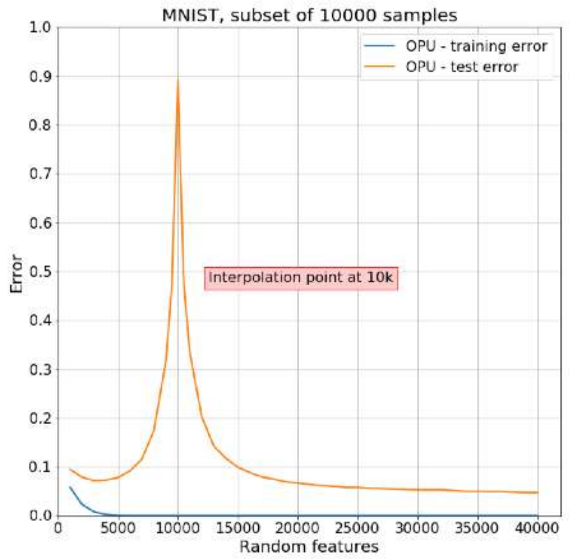
💪 OPUs have already been demonstrated in a diverse set of use cases:

✅ Molecular dynamics studies[9]



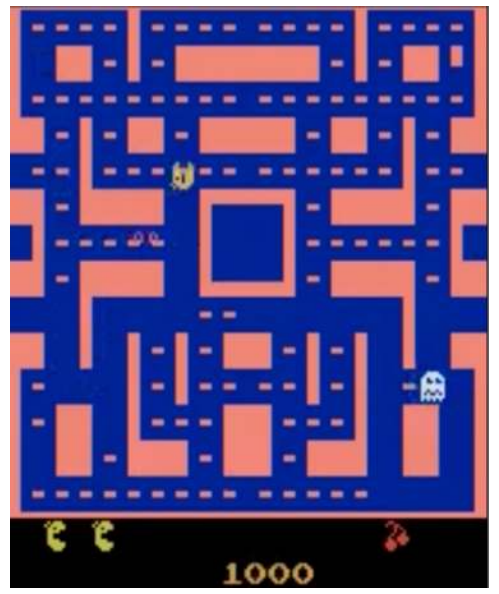
Anomaly detection on SARS-CoV-2 glycoprotein.

✅ Theoratical analysis of NNs[10]



Recovering the double descent curve.

✅ Reinforcement learning[11]



Playing PacMan with model-free RL.

⌨ Easy-to-use, the photonics are abstracted away: `opu.transform1d(x)` in Python; can use Numpy/PyTorch arrays.

LightOn CLOUD: the first and only photonic machine learning co-processor available in the cloud now!

Going linear with an holographic photonic core

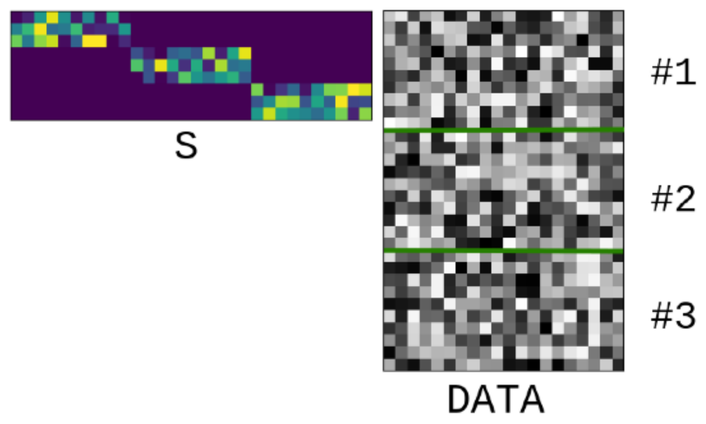
🤔 Current OPUs deliver a non-linear random projection, $|\mathbf{Bx}|^2$ not suitable for all applications.

🚀 We leverage holography to recover a linear operation from non-linear measurements.

✨ The magic: technology stack remains identical, enabling fast iterations.

💭 Massive potential for optical linear random projections:

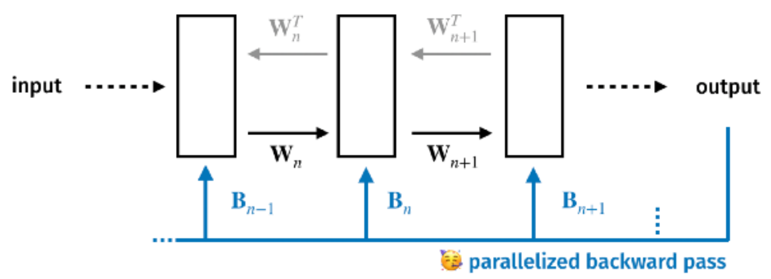
✅ Randomized linear algebra



Localized sketching to
compress large data streams.

in active development (this poster!)

✅ Optical training



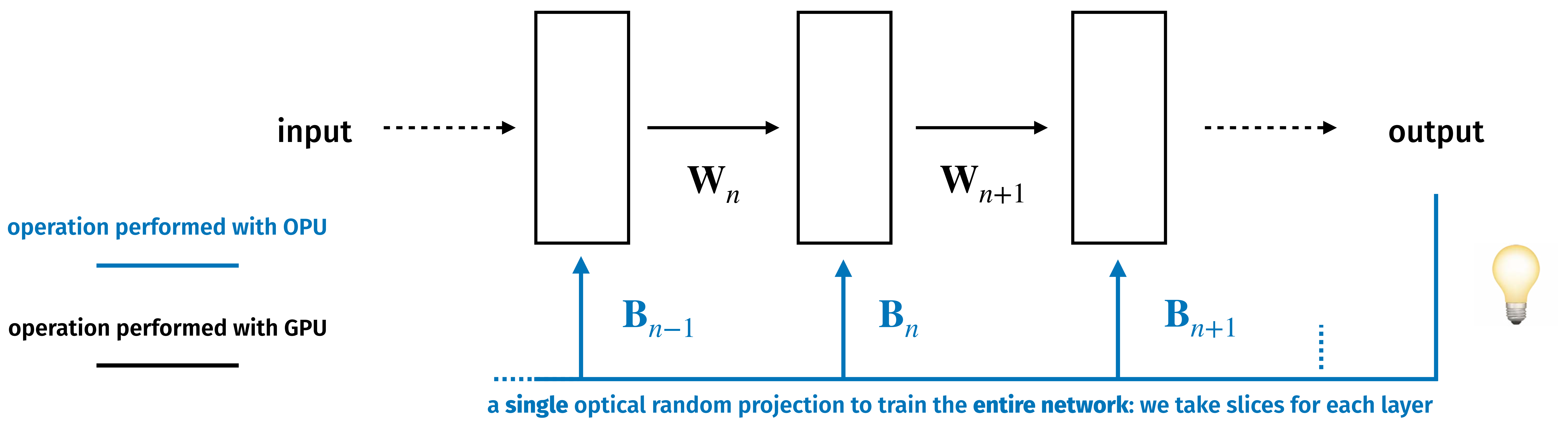
Optical Direct Feedback Alignment
to train neural networks.

🗓️ **Pre-release of holographic photonic core OPU in the cloud end of 2020.**

LightOn

Light-in-the-loop: photonic training of neural networks

💡 Implement the random projection of DFA optically:



🎂 **Agnostic to neural network architecture**: can be widely applied, beyond largest architectures in deep learning.

📈 **Demonstrated on MNIST**, with scaling to other tasks and architectures coming soon.
95.8% accuracy vs 97.7% on GPU for considered architecture

🎉 **The first time a neural network is trained with light-in-the-loop!**

Conclusion and outlooks

LightOn CLOUD: **OPUs are the first and only photonic machine learning co-processor <u>available in the cloud now</u>!**

More information at <u>cloud.lighton.ai</u>, including on our research program.

🎉 **The first time a neural network is trained with light-in-the-loop:**

- 🧠 We leverage learning beyond backpropagation to enable the use of advanced photonic hardware;
- ♾ Our accelerator is architecture-agnostic and scales to layers comprising millions of parameters.

🗓 **Pre-release of holographic photonic core OPU in the cloud end of 2020.**

Interested in knowing more about our technology?

<u>Check-out our white paper at lighton.ai!</u>

References

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[2] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[3] Alaa Saade, Francesco Caltagirone, Igor Carron, Laurent Daudet, Angélique Drémeau, Sylvain Gigan, and Florent Krzakala. Random projections through multiple optical scattering: Approximating kernels at the speed of light. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6215–6219. IEEE, 2016.

[4] Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In *Advances in neural information processing systems*, pages 1037–1045, 2016.

[5] Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback alignment scales to modern deep learning tasks and architectures. *arXiv preprint arXiv:2006.12878*, 2020.

[6] Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, pages 1–12, 2020.

[7] Sergey Bartunov, Adam Santoro, Blake Richards, Luke Marris, Geoffrey E Hinton, and Timothy Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems*, pages 9368–9378, 2018.

[8] Julien Launay, Iacopo Poli, and Florent Krzakala. Principled training of neural networks with direct feedback alignment. *arXiv preprint arXiv:1906.04554*, 2019.

[9] Amélie Chatelain, Giuseppe Luca Tommasone, Laurent Daudet, and Iacopo Poli. Online change point detection in molecular dynamics with optical random features. *arXiv preprint arXiv:2006.08697*, 2020.

[10] Alessandro Cappelli. Random projections did it again, April 2020.

[11] Martin Graive. Tackling reinforcement learning with the aurora opu, May 2020.

see https://medium.com/@LightOnIO