Light-in-the-loop: using a photonics co-processor for scalable training of neural networks

Julien Launay, Iacopo Poli, Kilian Müller, Igor Carron, Laurent Daudet, Florent Krzakala, Sylvain Gigan
LightOn
Paris, France
{julien, iacopo, kilian, igor, laurent, florent, sylvain}@lighton.ai

Abstract—As neural networks grow larger and more complex and data-hungry, training costs are skyrocketing. Especially when lifelong learning is necessary, such as in recommender systems or self-driving cars, this might soon become unsustainable. In this study, we present the first optical co-processor able to accelerate the training phase of digitally-implemented neural networks. We rely on direct feedback alignment as an alternative to backpropagation, and perform the error projection step optically. Leveraging the optical random projections delivered by our co-processor, we demonstrate its use to train a neural network for handwritten digits recognition.

I. INTRODUCTION

Deep neural networks are revolutionizing the way we approach computer vision, natural language processing, and even fundamental sciences. However, their efficiency relies on proper training, i.e. tuning the network’s weights. In the traditional supervised learning framework, training is performed on labeled data, whose amount has to scale with the complexity and size of the network. Traditionally, the training phase relies on the backpropagation algorithm [1]. However, this can become prohibitively expensive as networks grow in complexity; even more so when evaluating different configurations, combinations of hyperparameters (learning rate, dropout, etc.), or when retraining after new data samples are available. Whereas a number of hardware accelerators have been demonstrated for inference, building accelerators dedicated to a general-purpose algorithm such as backpropagation is challenging.

A number of alternative training methods have recently been proposed in the literature. As backpropagation prevents asynchronous processing of the layers of a neural network, since the update of a given layer depends on downstream quantities, there are valid reasons to find scalable alternatives. In particular, direct feedback alignment (DFA) [2] relies on a random projection of the model error to each layer as a training signal. Not only does this make the update of a layer independent of others, but it places a specific operation at the center of the training process: random projections [3].

The goal of this study is to show that DFA can be efficiently implemented using a photonic co-processor, built upon LightOn’s Optical Processing Unit [4]. This novel co-processor performs linear random projections fast, at large scale, and with low power consumption. It should be emphasized that this hybrid approach differs radically from an all-optical implementation of neural networks. We implement the forward path on traditional silicon-based digital chips, such as CPUs, GPUs or low-power alternatives; the photonic co-processor only helps in the feedback path for the training. Once training has been performed, the photonic co-processor is not required anymore for inference.
OBU modified to include off-axis holography. This optical co-
processor is capable of delivering linear random projections
at very large scales - up to more than a hundred billion
parameters. It is therefore well suited to be an accelerator for
neural network training with DFA, which we experimentally
demonstrate on the MNIST handwritten digits recognition task
[14].

II. METHODS

A. Direct feedback alignment

At layer $i$ of $N$, with $W_i$ its weight matrix, $b_i$ its biases,
$f_i$ its activation function, and $h_i$ its activations, the forward
pass of a neural network can be written as:

$$\forall i \in [1, \ldots, N]: a_i = W_i h_{i-1} + b_i, \ h_i = f_i(a_i) \ (1)$$

$h_0 = X$ is the input data and $h_N = f(a_N) = \hat{y}$ are the
predictions. With backpropagation, the update would be:

$$\delta W_i = -\frac{\partial L}{\partial W_i} = - [\langle W_{i+1} \delta a_{i+1} \rangle \odot f'_i(a_i)] h_{i-1}^T, \ \ (2)$$

with $\delta a_i = \frac{\partial L}{\partial a_i}$

With direct feedback alignment we have:

$$\delta W_i = - [\langle B \circ e \rangle \odot f'_i(a_i)] h_{i-1}^T \ (3)$$

B. Off-axis holography

In order to obtain a random projection of a given vector $e$
by a fixed matrix $B$ optically, the vector $e$ is first encoded onto a
coherent beam using a spatial light modulator. This beam then
propagates through a diffusive medium before the resulting
interference pattern (a speckle) is detected by a camera. Since
the camera detects the absolute square of the electromagnetic
field (the intensity), we record |Be|$^2$. Interfering with a reference
beam and the off-axis holography scheme we then recover the linear random projection Be.

III. RESULTS

We train a fully connected neural network using our optical
DFA training procedure. The network has two hidden layers
of 1024 units and uses tanh as the non-linearity. The error
vector $e$ is quantized to three values in order to be sent to the
input device of the optical system, using:

$$f(x) = \begin{cases} 
1 & \text{if } x > 0.1 \\
0 & \text{if } -0.1 < x < 0.1 \\
-1 & \text{if } x < -0.1 
\end{cases} \ (4)$$

The model is trained for 10 epochs, with ADAM [15], using
learning rate 0.01. With these parameters, we had a performance
of 95.8% test accuracy on MNIST. The same algorithm on a
GPU, with learning rate 0.001, reaches 97.6%, and the
algorithm without quantization of the error vector 97.7%.
The optical system runs at 1.5 kHz, with a maximal output size
of about $10^6$, that is it can perform 1500 random projections
of size $10^5$ per second, consuming about 30 W.

PERSPECTIVES

We have built and operated the first optical neural network
training accelerator. Our co-processor is architecture agnostic
and memory-less. For large-scale applications, it is competitive
with GPUs, and up to one order of magnitude more power
efficient.

By switching from the off-axis to a phase-shifting holography
scheme, it will be possible to scale input and output size up to $10^6$, and perform calculations involving more than
a trillion parameters. Future tests will involve scaling to even
larger networks or ensembles of networks.

We expect performance to improve with the optimization
of the currently available components, as well as as with the
development of future components. A better understanding of
DFA will also help widen the scope of applications of this
accelerator.

REFERENCES

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal
representations by error propagation,” tech. rep., California Univ San
Diego La Jolla Inst for Cognitive Science, 1985.
[2] A. Nokland, “Direct feedback alignment provides learning in deep
neural networks,” in Advances in neural information processing systems,
pp. 1037–1045, 2016.
[3] S. Dasgupta, “Experiments with random projection,” arXiv preprint
arXiv:1301.3849, 2013.
[4] A. Saade, F. Calagirone, I. Carron, L. Daudet, A. Drémeau, S. Gigan,
and F. Krzakala, “Random projections through multiple optical scat-
tering: Approximating kernels at the speed of light,” in 2016 IEEE
International Conference on Acoustics, Speech and Signal Processing
(ICASSP), pp. 6215–6219, IEEE, 2016.
[5] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa,
S. Bates, S. Bhatia, N. Boden, A. Borchers, et al., “In-datacenter
performance analysis of a tensor processing unit,” in Proceedings of the
44th Annual International Symposium on Computer Architecture, pp.
1–12, 2017.
[6] Z. Jia, B. Tillman, M. Maggoni, and D. P. Scarpazza, “Dissecting
the graphcore ipu architecture via microbenchmarking,” arXiv preprint
arXiv:1912.03413, 2019.
[7] M. Smelyanskiy, “Zion: Facebook next-generation large memory train-
ing platform,” in 2019 IEEE Hot Chips 31 Symposium (HCS), pp. 1–22,
IEEE, 2019.
[8] “Pytorch at tesla - andrej karpathy, tesla.” https://www.youtube.com/
watch?v=OBlh7XKXIDE. Accessed: 2020-03-19.
[9] D. Han, J. Lee, J. Lee, and H.-J. Yoo, “A low-power deep neural network
online learning processor for real-time object tracking application,” IEEE
Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 5,
pp. 1794–1804, 2018.
[10] D. Han, J. Lee, J. Lee, and H.-J. Yoo, “A 1.32 tops/w energy efficient
deep neural network learning processor with direct feedback alignment
based heterogeneous core architecture,” in 2019 Symposium on VLSI
Circuits, pp. C304–C305, IEEE, 2019.
[11] B. Crafton, M. West, P. Basnet, E. Vogel, and A. Raychowdhury, “Local
learning in ram neural networks with sparse direct feedback alignment,”
in 2019 IEEE/ACM International Symposium on Low Power Electronics
and Design (ISLPED), pp. 1–6, IEEE, 2019.
[12] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, “Training of photonic
neural networks through in situ backpropagation and gradient measure-
ment,” Optica, vol. 5, no. 7, pp. 864–871, 2018.
[13] X. Guo, T. D. Barrett, Z. M. Wang, and A. Lvovsky, “End-to-end
optical backpropagation for training neural networks,” arXiv preprint
arXiv:1912.12256, 2019.
[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning
applied to document recognition,” Proceedings of the IEEE, vol. 86,
no. 5, pp. 2278–2324, 1998.
[15] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,”
arXiv preprint arXiv:1412.6980, 2014.