Accelerating deep learning with high energy efficiency: From microchip to physical systems

Huanhao Li,[1,2] Zhipeng Yu,[1,2] Qi Zhao,[1,2] Tianting Zhong,[1,2] and Puxiang Lai[1,2,3,*]

[1]Department of Biomedical Engineering, Hong Kong Polytechnic University, Hong Kong, China

[2]Shenzhen Research Institute, Hong Kong Polytechnic University, Shenzhen 518057, China

[3]Photonics Research Institute, Hong Kong Polytechnic University, Hong Kong, China

*Correspondence: puxiang.lai@polyu.edu.hk

Received: March 17, 2022; Accepted: April 23, 2022; Published Online: April 26, 2022; https://doi.org/10.1016/j.xinn.2022.100252

Citation: Li H., Yu Z., Zhao Q., Zhong T., and Lai P. (2022). Accelerating deep learning with high energy efficiency: From microchip to physical systems. The Innovation **3(4)**, 100252.

In the era of digits and internet, massive data have been continuously generated from a variety of sources, including video, photo, audio, text, internet of things, etc. It is intuitive that more accurate patterns can be obtained by feeding more data for effective analysis; despite the data redundancy, a clearer picture can be delineated for better decision-making. However, traditional methods, even in machine learning, do not benefit from the expanding amount of data, whose performance nearly saturates when the data collection is large enough (Figure 1A). Such a dilemma emerges due to their limited capability and insufficient supply of computation power in the past.

A breakthrough was pinned in 2012 with deep learning: a deep neural network (DNN) can be effectively trained on graphics-processing units (GPUs) of phenomenal performance.[1] The DNN mimics the biological neural networks in brains with layer-stacked transformations of sufficient complexity to approximate arbitrary functions. Major mathematical operations in DNNs, such as matrix multiplication, convolution, and other customized repetitive computations, can be engineered into an in-parallel configuration, which matches well with the computational mechanism of GPUs. Processing efficiency is therefore boosted and outperforms previous realizations done with serial computation units (e.g., central-processing units [CPUs]). Deep learning has inspired a broad range of applications in computer vision (CV), natural language processing (NLP), biomedicine, games, and many others. Computational requirements from these fast-expanding applications of DNNs, on the other hand, significantly outstrip the development of the chips on silicon (i.e., Moore's law). The current computational facilities of deep learning are accompanied with huge energy costs. Therefore, DNN accelerators are urgently demanded for fast processing and energy efficiency.

The learning accelerators can be either general- or specific purpose, with implementations based on microchip/conventional electronics (digital) or physical systems (analog) by tackling the notorious "memory wall" and "power wall." Orders-of-magnitude improvements have been achieved regarding both processing speed and energy efficiency. These achievements significantly shorten the implementations of large DNNs, even fed with millions or trillions of data. For now, an accelerated DNN for image classification or a language translator can be trained

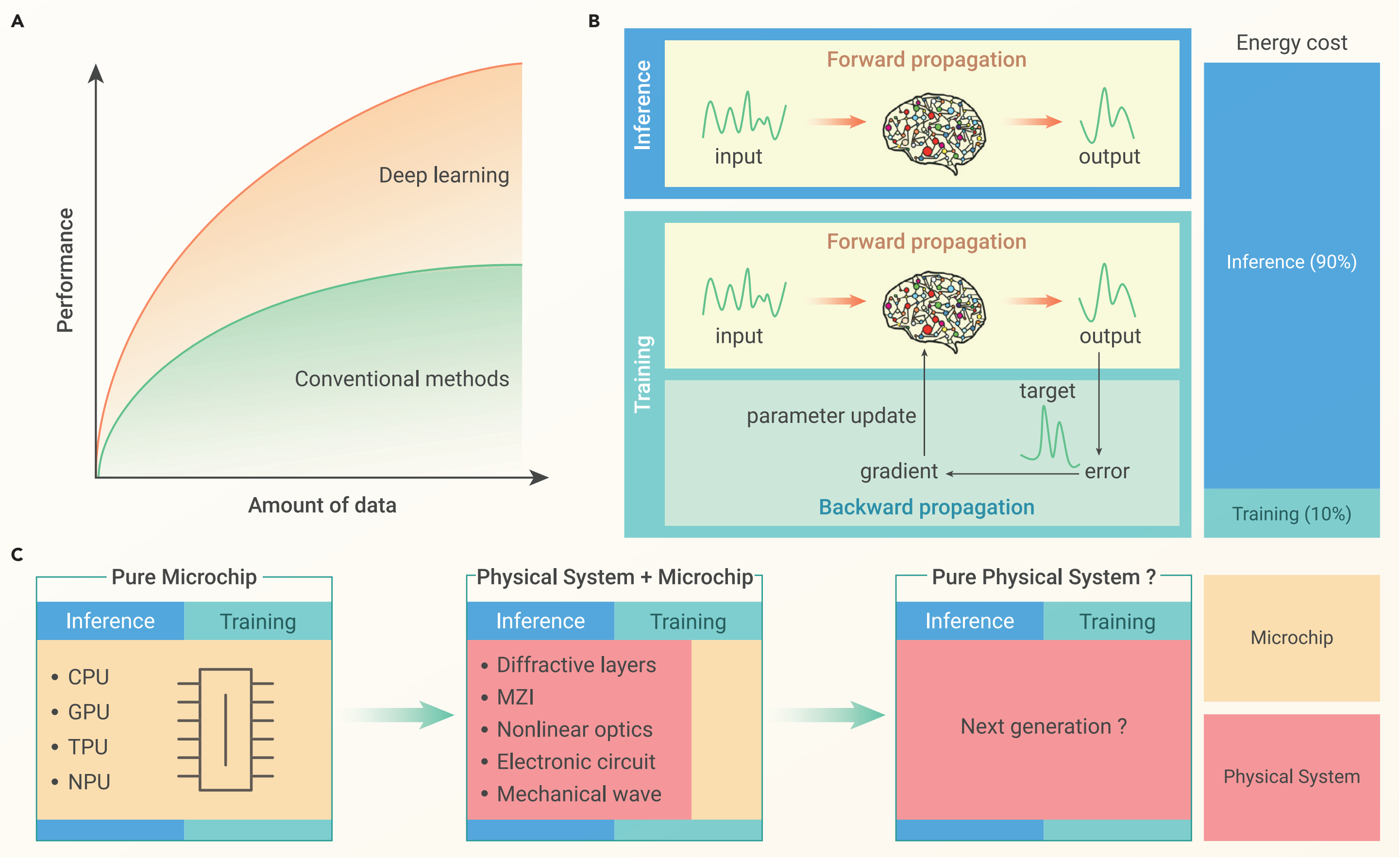

Figure 1. Diagram of deep learning realizations (A) Comparisons between deep-learning and conventional methods on how the amount of the data affects the performance. (B) Procedures (inference, backward propagation including error/gradient calculation and parameter update) contained in deep learning could be accelerated. The side bar denotes the energy cost for the inference (90%) and training (10%) phases. (C) Development trend of the DNN accelerator from pure microchip to physical system. MZI, Mach-Zehnder interferometers.

to work within a couple of hours; without these accelerators, the same task may take several weeks. Such convincing efficiency has made these artificial intelligence (AI)-specific technologies more accessible and significantly accelerates the transfer of deep-learning-based products toward commercialization.

Two essential phases, namely training and inference (Figure 1B), are targeted to speed up the learning process. Different requirements pose for these two phases. The training phase involves the feeding of massive data into the DNN for designed tasks to update and finalize the parameters of the DNN via backward propagation optimization. After training, the DNN with fixed (or optimized) parameters comes into the inference phase to predict the desired result with the input data. In many applications, training is a one-time effort, but the functioning of the trained DNN in inference occupies significantly longer durations. Hence, inference consumes up to 90% of the energy cost in deep learning,[2] which is the very concern for most accelerators.

In inference, most of the processing time is occupied by matrix multiplication, whose mathematical isomorphism has been devised for hardware acceleration. The most widely used and easiest-to-access chips are the general GPUs, which can perform at least an order-of-magnitude faster when Tensor Cores are customized by Nvidia to boost the matrix-multiplication efficiency. Moreover, to further improve the energy efficiency, DNN-specific computing units are developed, such as the tensor-processing units (TPUs) announced by Google. A TPU is designed for its own deep-learning framework (TensorFlow) and neural-processing units (NPUs), which “hardwarelizes” the architecture of the DNN on the chips. Because DNN computations in these highly specific customized configurations are not interrupted by other tasks, energy efficiency can be improved by several folds for the same computational tasks.

Physical systems provide analog solutions to mimic the above-mentioned neural processing and computation, in which data transformation is dictated by physical mechanisms and requires negligible power consumption. This forms a sharp contrast to microchip systems, where digital and/or Boolean operations of high precision and low noise ensure computational accuracy. Deep learning is able to deal with data of low precision and high noise, which usually arises in physical systems. Therefore, noise can be conquered, to some extent, by introducing noise in the training phase. And, to be more rigorous, system stability can be physically strengthened by adding stabilizer into the packing of the system, though sophisticated engineering is necessary. Optics (or photonics), for example, is of particular interest since the light-matter interaction essentially companies Fourier transforms, matrix multiplication, convolution, and many other operations such as nonlinearity.[3] An optical system is therefore inherently empowered to fulfill the tasks in CV, such as image classification and transformation, at the speed of light. To achieve multiple convolutional kernels, an optical component, enabling modulation in amplitude and/or phase in multiple channels, needs to be embedded into the system. This criterion matches well with optical components, such as ground glass diffusers, diffractive layers, multimode fibers, and specifically designed metasurfaces. The weights of the DNN kernels represented by these components are fixed with certain physical configurations (e.g., customized three-dimensional [3D]-printed layers[4]) based on the training simulated in a computer. Free-space optics therefore focuses on the inference phase of deep learning. With proper manipulation, an optical system becomes versatile in deep learning, with applications not limited to CV but also NLPs such as vowel classification. In addition to mimicking mathematical operations in DNNs by the physical system, it is also feasible to train the hardware’s physical transformation to match specific task/applications, such as ultrafast optical second-harmonic generation, multimode mechanical oscillations, and the analog dynamics of a nonlinear triode. Thanks to continuous development, the inference performance for the same task is approaching the state-of-art results provided by microchip systems.

Another approach to extend the applications of the physical system lies in the training phase, which accounts for the rest of the energy cost in deep learning ($\sim$10%). Physical systems are not commonly designed to speed up the training phase because their configurations are usually passively determined and the embedded DNN parameters are not trainable (or changeable) without re-configuration. The poor flexibility poses obstacles to fine-tune the physical networks for better performance and transfer the function of system to other applications. In neuromorphic computing, tunable units (e.g., phase-change components in photonics circuits) are designed to store the weights of the synapse, accompanied by control circuits.[5] Trainable parameters can also be encoded or combined with the input data; the physical systems here merely serve as an operator (e.g., adders, multipliers, or nonlinear transformations) actuated by physical mechanisms. More presentations of the physical systems are expected to inspire and incubate more AI-specific edge modules or smart sensors, especially for scenarios where low latency is necessary. The internet of things (IoT), like self-driving cars, smart home systems, etc., could be more beneficial as it inherently processes physical data from the environment and desires fast processing.

Achieving trainable physical systems is not the end goal (Figure 1C). Microchips are inevitable accessories for analog-to-digital/digital-to-analog conversion (ADC/DAC), data processing, data storage, and workflow control. Heat dissipation and temporal latency of those electronics could loop back as the bottleneck of physical systems in deep learning. Seamless interactions between the physical mechanism and the accessories are expected, which necessitates specific design and sophisticated engineering for their integration. Furthermore, co-packing and miniaturization for them is also a critical challenge since not all physical systems can be engineered and manufactured in small size. These bulky systems limit the number of neurons in a single network and, hence, the network’s complexity. Currently, the largest number of neurons in a reported trainable physical system in a chip-based realization occurs in the photonics circuit (on the scale of 10^2), while the numbers in large DNNs, like ResNet101 and Unet, are on the scale of 10^5. Such restrictions are expected to be addressed so that the ability of AI-specific physical systems can be equivalent to the DNNs running on computers, which will enable considerably more complex tasks, such as image segmentation, image style transformation, language translation, etc.

In summary, various accelerators have been developed to improve DNN computation regarding processing speed as well as energy efficiency. An overall tendency of the computation platforms has been seen, evolving from conventional electronics to physical systems, with progressive reductions in energy consumption. While there are limitations for current realizations of physical systems, vigorous development is strongly looked forward to to continuously improve energy efficiency with the desired performance.

REFERENCES

1. Alex Krizhecsky, I.S., and Hinton, Geoffrey E. (2012). ImageNet classification with deep convolutional neural network. Proced. Technol. **15**, 474–483.
2. Patterson, D., Gonzalez, J., Le, Q., et al. (2021). Carbon Emissions and Large Neural Network Training. Preprint at arXiv. https://doi.org/10.48550/arXiv.2104.10350.
3. Wetzstein, G., Ozcan, A., Gigan, S., et al. (2020). Inference in artificial intelligence with deep optics and photonics. Nature **588**, 39–47.
4. Lin, X., Rivenson, Y., Yardimci, N.T., et al. (2018). All-optical machine learning using diffractive deep neural networks. Science **361**, 1004–1008.
5. Shastri, B.J., Tait, A.N., Ferreira de Lima, T., et al. (2021). Photonics for artificial intelligence and neuromorphic computing. Nat. Photon. **15**, 102–114.

ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China (NSFC) (81930048, 81627805), the Hong Kong Research Grant Council (15217721, R5029–19, and C7074-21GF), the Hong Kong Innovation and Technology Commission (GHP/043/19SZ and GHP/044/19GD), and the Guangdong Science and Technology Commission (2019A1515011374 and 2019BT02X105).

DECLARATION OF INTERESTS

The authors declare no competing interests.