Conference materials
UDC 535.8
DOI: https://doi.org/10.18721/JPM.163.213

Features of the construction photonic tensor cores for neural networks

N.I. Popovskiy[1]✉, V.V. Davydov[1, 2], V.Yu. Rud [3, 4]

[1] The Bonch-Bruevich St. Petersburg State University of Telecommunications, St. Petersburg, Russia;

[2] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia;

[3] Ioffe Institute, St. Petersburg, Russia;

[4]All-Russian Research Institute of Phytopathology, Moscow Region, Russia

✉ nikitanikita24@mail.ru

Abstract. The demand for efficient and high-performance computing systems has led to the development of photonic-based technologies for machine learning. One of the key components of these systems is the photonic tensor core, which performs matrix operations at high speed and low power consumption. In this article, we review the features of photonic tensor cores and their construction for use in neural networks. We discuss the advantages of photonic-based technologies over traditional electronic-based systems, as well as the challenges in their implementation. We also highlight recent advancements in the development of photonic tensor cores for machine learning applications.

Keywords: photonic tensor cores, neural networks, optical computing, photonics, machine learning, deep learning, data processing

Citation: Popovskiy N.I., Davydov V.V., Rud V.Yu., Features of the construction photonic tensor cores for neural networks, St. Petersburg State Polytechnical University Journal. Physics and Mathematics. 16 (3.2) (2023) 81–86. DOI: https://doi.org/10.18721/JPM.163.213

Материалы конференции
УДК 535.8
DOI: https://doi.org/10.18721/JPM.163.213

Особенности построения фотонных тензорных ядер для обучения нейронных сетей

Н.И. Поповский[1]✉, В.В. Давыдов[1,2], В.Ю. Рудь[3,4]

[1] Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича, Санкт-Петербург, Россия;

[2] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия;

[3] Физико-технический институт им. А.Ф. Иоффе РАН, Санкт-Петербург, Россия;

[4] Всероссийский научно-исследовательский институт фитопатологии, Московская область, Россия

✉ nikitanikita24@mail.ru

Аннотация. Спрос на эффективные и высокопроизводительные вычислительные системы привел к разработке основанных на фотонике технологий машинного обучения. Одним из ключевых компонентов этих систем является фотонное тензорное ядро, которое выполняет матричные операции с высокой скоростью и низким энергопотреблением. В этой статье мы рассмотрим особенности фотонных тензорных ядер и их конструкцию для использования в нейронных сетях. Мы обсуждаем преимущества технологий, основанных на фотонике, перед традиционными электронными системами, а также проблемы, связанные с их внедрением. Мы также подчеркиваем недавние достижения в разработке фотонных тензорных ядер для приложений машинного обучения.

Ключевые слова: фотонные тензорные ядра, нейронные сети, оптические вычисления, фотоника, машинное обучение, глубокое обучение, обработка данных

Ссылка при цитировании: Поповский Н.И., Давыдов В.В., Рудь В.Ю. Особенности построения фотонных тензорных ядер для обучения нейронных сетей // Научно-технические ведомости СПбГПУ. Физико-математические науки. 2023. Т. 16. № 3.2. С. 81–86. DOI: https://doi.org/10.18721/JPM.163.213

Introduction

The challenge of emulating brain functions continues to fascinate and inspire human ingenuity, and has also proven to be of practical value to modern societies. One approach that has gained popularity in Artificial Intelligence (AI) is Machine Learning (ML) through the use of neural networks (NN). This involves training a system to autonomously classify and make decisions about new data, and once trained, the NN can be utilized to recognize and categorize patterns and objects. This is applied in various areas of science and technology, especially those related to space systems [1–5].

Neural networks (NNs) typically consist of multiple layers of interconnected neurons or nodes, and the configuration of each layer, as well as the interconnectivity of the network as a whole, is crucial for the network's ability to perform its intended task. The processing within a NN's connected layers relies heavily on vector matrix math operations, involving the multiplication of large matrices of input data and weights based on the training data. As the complexity and depth of NNs increase, their ability to perform these large matrix multiplications efficiently and quickly requires substantial bandwidth and low latency. This is extremely important in many cases. For example, for fiber-optic communication lines (FOCL) in systems of radar stations, when it is necessary to accompany a large number of objects, these functions are in great demand [6–8].

Since the beginning of the computing era, researchers have been exploring efficient methods to multiply matrices due to its ubiquity in various applications, including neuromorphic computing. Developing a platform that can perform matrix multiplication faster and more energy-efficiently is crucial in solving linear algebraic problems like inverting matrices, solving linear equations, and finding determinants. In fact, even fundamental graph algorithms can be hindered by slow matrix multiplication. This creates a number of problems in information transmission systems with intelligent processing [4, 9–11].

Matrix operations on a general-purpose processor are performed serially and require continuous access to cache memory, which creates a bottleneck known as the "von Neumann bottleneck". Specialized architectures, such as Graphic Process Units (GPUs) and Tensor Process Units (TPUs), have been developed to reduce this bottleneck and enable advanced machine learning models. These architectures are designed with domain-specific optimizations, such as parallel processing for convolutions or Matrix-Vector Multiplications (MVM), allowing for the deployment of systolic algorithms unlike CPUs [2, 9, 12, 13].

The advantages of using electromagnetic signals may be limited due to the need for conversion by optoelectronic and electro-optical methods, as well as repeated access to digital and non-volatile memory, which can lead to slower operation and high energy consumption. In this regard, the use of heterogeneously integrated optimized photonic memory, which can store information in a non-volatile state, is a great advantage, especially for projects using neural networks, where weights are rarely updated.

Method of constructing photonic tensor cores

To achieve this functionality, a multi-state photonic memory device has been developed, in which a set of cells are located between two resonant rings to select the appropriate wavelength at the input and output. Once the memory states are set in this photonic core, it is possible to perform the calculation functions completely passively. Selective recording is achieved by changing the phase of a certain number of cells that have been deposited on the waveguides by local

electrostatic heating, which leads to crystallization or amorphization and, accordingly, a change in the modal refractive index of the waveguide in a reversible process.

The dot product mechanism multiplies the i-th row of the input matrix A by the *j*th column of the kernel B. The input matrix is represented in Fig. 1 by WDM signals, which are modulated using high-speed modulators such as Mach-Zehnder [14]. The column of the core matrix is loaded into the photonic memory with a given weight state. The interaction of light matter with memory leads to a change in the phase of the input signals, which are spectrally filtered by micro-ring resonators, weighed using amplitude modulation and subjected to element-wise multiplication. The resulting products of the elements are summed using a photodetector that performs the MAX (Dij) operation. Quantization is used in the electrical absorption circuit.

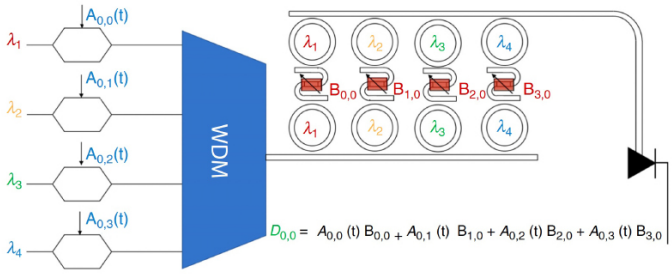

Fig. 1. Block diagram of the photonic tensor cores

In our approach to the implementation of photonic neural networks, micro-rings are used only for passive selection of the frequency that will be modulated by photonic storage devices, unlike other implementations that use actively tuned micro-ring modulators for filtering [14, 15]. This allows us to control inter-channel crosstalk more precisely and potentially increases the number of wavelengths in the multiplexing scheme with dense wavelength separation without affecting the absorption coefficient and the associated fluctuations in the quality coefficient. In addition, our architecture includes programmable photonic storage devices with a low loss level and multiple states for each pair of micro-ring resonators, which are able to store information without static power consumption and do not significantly contribute to total losses.

The $Ge_2Sb_2Se_5$ material was chosen for the implementation of photonic memory cores, since in the amorphous state it has a wide area of transparency for telecommunication wavelengths and can be used to create high-performance non-volatile photonic storage devices with multiple states. This material has very low optical absorption, which makes it promising for multi-state devices and avoids the use of high-power lasers and extremely low noise detectors. To set the extracted weights, we use electrothermal switching, which reversibly records each memory state by selective transition between amorphous and crystalline phases. To do this, heat is supplied to the material from the outside using Joule heating, which is activated by successively supplying various pulses to the cell through connected devices.

To minimize losses, the choice of material and the location of the electrodes in relation to the waveguide were specially developed, which allowed the use of metal with excellent thermal properties and low optical losses. Adjustment of the frequency and intensity of electrical pulses applied to tungsten electrodes is necessary to provide the necessary thermal energy for phase switching in $Ge_2Sb_2Se_5$. To create an effective resistive heater that will not create losses, it is possible to use doped silicon, silicide, indium-tin oxide or graphene electrodes that will be located next to the waveguide. The change in the absorption coefficient during the phase transition is investigated using light signals associated with the memory of the phase transition.

During the network training process, the weights are derived by employing electrothermal switching of individual states of photonic memories, as opposed to the previously used optical pulses. This technique involves writing each memory state reversibly by selectively transitioning between amorphous and crystalline phases through electrothermal switching induced by Joule heating. In our approach, external heat is applied to the material using joule heating of a tungsten metal layer in direct contact with the wire. Different pulse train profiles, based on the type of transition required, are applied to the wire through connections in series to the device.

Results and Discussion

Our photonic memories consist of PCM (Phase-Change Memory) wires arranged in a grating pattern. Each wire represents a quantized state, and we use 30-nm-thin and 250-nm-wide reprogrammable PCM-wires. By considering the condition where all wires are in the amorphous state as the highest state, we can achieve a 4-bit memory for each element of the kernel (Bij) using just 15 reprogrammable wires. The total length of this memory is only 8 micrometers, not including the electrical circuitry.

The insertion loss, defined as the decrease in optical power transmitted when some wires are switched to the crystalline state, is approximately 1 dB for the 4-bit multilevel memory. This results in discrete power levels for each quantized state. When all wires are in the amorphous state, the transmitted optical power is denoted as P_0.

In this configuration of photonic memory, uniform quantization is achieved, where each state corresponds to one quantization step. In 4-bit photon memory, the quantization step is 0.2 dB per state, and the maximum attenuation coefficient is approximately 3.5 dB, as shown in Fig. 2. The attenuation coefficient is calculated by dividing the optical power transmitted in two extreme configurations, that is, when all wires are in the crystalline state and when all wires are in the amorphous state.

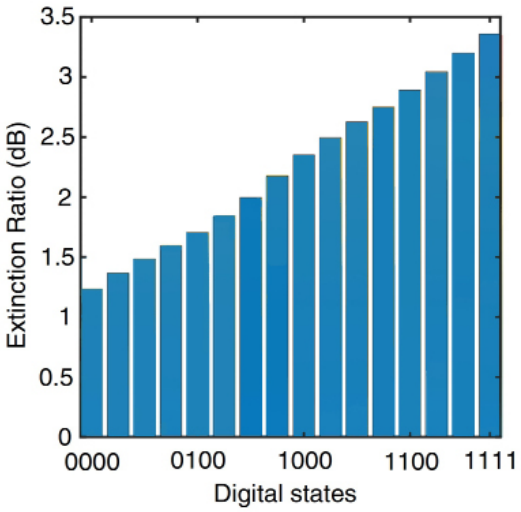

Fig. 2. Extinction ratio (ER) for a 4-bit photonic memory as a function of digital states, for an increased number of crystalline wires, the ER increases linearly and uniformly

Instead of using separate wires for each state, a smaller number of films of different lengths can be used to generate the remaining states by recording these films in different combinations. When using the ratio of linear losses per unit length, it is possible to implement 4-bit memory using only four films of variable length, where the losses in the crystal state correspond to the states 1000, 0100, 0010 and 0001. This binary weighing approach reduces the number of heaters and tungsten contact pads while maintaining the total installation area. However, each suspended state will require different write/erase time and voltage, which requires further optimization.

Although it is important to note that when performing logical inference, due to the stability of the NNs achieved through timely training, low-bit quantization of weights is also possible, which allows you to get a really efficient and accurate output for quantized weights with low resolution. If the system is going to be used to perform relatively simple inference tasks at the edge of the network, it may not require high resolution.

Thus, we propose a tensor core module implemented in photonics, which relies on photonic multiplexing (WDM) signals weighted after filtering using a constructed multi-component photonic memory based on $Ge_2Sb_2Se_5$ cells formed on a waveguide. Photonic memory is reprogrammed by selectively changing the phase (amorphous/crystalline) of wires using electrothermal switching by Joule heating induced by tungsten electrodes. If necessary, the programming of the photonic memory can be implemented in parallel, if necessary, or, alternatively, this photonic

tensor core can work as a passive system with a given core matrix; that is, there will be neither dynamic nor static power dissipation. Another key technological feature of this design is that photonic memory does not introduce additional losses, which avoids repeaters and optical amplifiers, cumbersome intersections and transformations between electrical and optical domains. The architecture shows the execution time limited only by the photon flight time in the chip, which is a function of the number of wavelengths and the delay of the photodetector after programming the core matrix and processing optical input data. The simultaneous development of new materials and the development of integrated photonic memory can allow the implementation of mechanisms based on the proposed scheme, capable of inherently performing accumulation and matrix multiplication with floating point, and, consequently, open the way to the implementation of fully optical photonic tensor blocks, which can significantly accelerate the execution of intelligent tasks at the network boundary, without requiring electro-optical transformations and access to external memory.

Conclusion

The proposed Photonic Tensor Core (PTC) operates based on the outlined scheme and is capable of passive matrix multiplication with 4-bit precision. This operation only needs to occur once, during the storage of weights in the photonic network. The PTC functions independently of any logical architecture and doesn't necessitate data transduction from external memory for inference. This characteristic positions it as a comprehensive analog processor, akin to recently developed counterparts. Specifically, during inference tasks, our architecture conducts tensor operations with a time complexity of O(1), and its static power consumption approaches negligible levels. This efficiency arises from the system acting as a passive filter, relying solely on light-matter interactions with pre-stored states in the photonic memory. Unlike logic operations that require optical switching, our system operates by leveraging these interactions and accessing inputs directly from the optical domain. The photonic memory retains previously saved kernels, assuming they were stored during a prior instance, and the inputs are readily accessible from the network's edge.

REFERENCES

1. **Mihov E.D., Nepomnyashchiy O.V.,** Selecting informative variables in the identification problem Journal of Siberian Federal University - Mathematics and Physics. 9 (4) (2016) 473–480.

2. **Lin X., Rivenson Y., Yardimci N.T., Veli M., Luo Y., Jarrahi M., Ozcan A.**, Science 361, 1004 (2018).

3. **Miscuglio M., Mehrabian A., Hu Z., Azzam S.I., George J., Kildishev A.V., Pelton M., Sorger V.J.**, Opt. Mater. Express 8, 3851 (2018).

4. **Ryzhenko I.N., Lutsenko A.E., Varygin O.G., Nepomnyashchiy O.V.,** Carrier compensation mode implementation in satellite communication channels. In: Proceedings of 2019 International Siberian Conference on Control and Communications, SIBCON 2019, vol. 8729665 (2019) 23–29.

5. **Shen Y., Harris N. C., Skirlo S., Prabhu M., Baehr-Jones T., Hochberg M., Sun X., Zhao S., Larochelle H., Englund D., Soljačicr M.,** Nat. Photonics 11, 441 (2017).

6. **Miscuglio M., Mehrabian A., Hu Z., Azzam S.I., George J., Kildishev A. V., Pelton M., Sorger V.J.**, Opt. Mater. Express 8, 3851 (2018).

7. **Popovskiy N.I., Davydov V.V., Rud V.Y.,** Features of construction of fiber-optic communication lines with orthogonal frequency-division multiplexing, St. Petersburg Polytechnic University Journal. Physics and Mathematics. 15(3.2) (2022) 178–183.

8. **Krivenko Y.E., Logunov S.E., Davydov M.N., Andreeva E.I., Andreev D.P.,** Selection Features of Different Standard Optical Fiber in CCTV Fiber-Optics Systems, Springer Proceedings in Physics. 268 (2022) 539–543.

9. **Popovskiy N.I., Davydov V.V., Rud V.Yu.,** Features of the construction of photonic integrated circuits for communication systems. Journal of Physics: Conference Series 2086(1) (2021) 012163.

10. **Filatov D.I., Galichina A.A., Vysoczky M.G., Yalunina T.R.,** Features of transmission at analog intermediate frequency signals on fiber – optical communication lines in radar station, Journal of Physics: Conference Series. 917(8) (2017) 082005.

11. **Reznikov B., Rodin S., Popovskiy N., Isaenko D., Stepanenkov G., Vakorina D.,** Development to High-Rate Fiber Optic Communication Line with Orthogonal Frequency-Division Multiplexing. 2022 International Conference on Electrical Engineering and Photonics (EExPolytech) (2022).

12. **Podstrigaev A.S., Lukiyanov A.S., Smolyakov A.V., Nikitina M.I.,** The expediency of fiber-optical communication line used in different schemes of receiver tract of the radio-monitoring complex, Journal of Physics: Conference Series. 1368(2) (2019) 022027.

13. **Tarasenko M.Y., Lenets V.A., Akulich N.V., Yalunina T.R.,** Features of use direct and external modulation in fiber optical simulators of a false target for testing radar station, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 10531 LNCS (2017) 227–232.

14. **Isaenko D., Reznikov B., Rodin S., Stepanenkov G., Popovskiy N., Vakorina D.,** Development of Communication Channel for data Transmission Over Single-Mode Optical Fiber in Environmental Monitoring System from Remote Multifunctional Complexes. 2022 International Conference on Electrical Engineering and Photonics (EExPolytech) (2022).

15.**Miscuglio M.,** Photonic tensor cores for machine learning, Applied Physics Reviews. 7(2020) 031404-1-031404-10.

THE AUTHORS

POPOVSKIY Nikita I.
nikitanikita24@mail.ru
ORCID: 0000-0003-3457-0370

RUD Vasiliy Yu.
rudvas.spb@gmail.com
ORCID: 0000-0002-5731-0849

DAVYDOV Vadim V.
davydov_vadim66@mail.ru
ORCID: 0000-0001-9530-4805

Received 25.07.2023. Approved after reviewing 10.08.2023. Accepted 10.08.2023.