Optical Convolutional Neural Networks

Subjects: Computer Science, Hardware & Architecture

Contributor: Ming Li, Nuannuan Shi, Xiangyan Meng, Guangyi Li, Wei Li, Ninghua Zhu

As a leading branch of deep learning, the convolutional neural network (CNN) is inspired by the natural visual perceptron mechanism of living things, showing great application in image recognition, language processing, and other fields. Photonics technology provides a new route for intelligent signal processing with the dramatic potential of its ultralarge bandwidth and ultralow power consumption, which automatically completes the computing process after the signal propagates through the processor with an analog computing architecture.

Keywords: convolutional neural networks ; optical computing ; photonics signal processing

1. Introduction

Convolutional neural networks (CNNs), as an important category of deep neural networks, are inspired by the natural visual perceptron mechanism of living things [1]. Since the first modern sense framework of CNNs, known as LeNet-1 [2], emerged in 1989, numerous representative CNN frameworks have been developed, including LeNet-5 (1998) [3], AlexNet (2012) [4], ZFNet (2014) [5], VGGNet (2015) [6], GoogLeNet (2015) [7], and ResNet (2016) [8]. Meanwhile, abundant progress has been made to deepen CNNs' complexity and reduce the number of parameters [9][10][11]. Owing to the continuous optimization of network frames, CNNs have been widely used in image recognition [2][3][4][5][6][7][8][9][11][12][13], speech recognition [14][15][16], gaming [17][18], medicine [19][20], autonomous driving [21][22], and other fields.

The explosive increase in Internet data year by year has called for more intelligent and effective data processing [23]. As is well-known, there is a positive correlation between the accuracy of a CNN and the number of parameters [24]. Therefore, it has more stringent requirements on the computing hardware due to the demands of massive data processing and high-precision processing. For electrical hardware processors, performance improvements have followed Moore's Law over the past few decades [25][26]. As the chip manufacturing process has gradually approached its physical limitations in recent years, the growth rate of single-chip computing power has gradually slowed [27][28], and semiconductor technology has entered the post-Moore era. Additionally, using the Von Neumann computing paradigm in the traditional computing hardware, such as CPU, GPU, FPGA, ASIC, etc., it is an indisputable fact that the discrete architecture of processor and memory makes it inevitable to trade-off between bandwidth and power consumption [29][30][31][32]. Hence, it is an obvious sharp conflict with the ever-increasing demand for high-performance processing and the slowing growth of computing power [28].

Optical devices, as an alternative, have been regarded as a competitive candidate in the "more than Moore" era [33] with the superiority of ultralarge bandwidth and ultralow power consumption. Compared with electrical vector-matrix multiplication (VMM), it is able to achieve better performance using optical devices, with the computing speed increasing by three orders of magnitude and the power consumption decreasing by two orders of magnitude [34]. In recent years, optical computing solutions, because of their intrinsic high computing speed [35], high computational density [36][37], and low power consumption [38][39], have been massively demonstrated by means of both discrete systems and integrated chips. CNNs, as one of the main branches, require more than 80% of full calculations to execute the convolution operation [40]. Accelerating the convolution process in the optical domain provides a subversive way to improve the computing speed and decrease the power consumption.

2. Development of Optical Convolution Neural Network

Generally, a CNN is composed of convolutional layers, pooling layers, fully connected layers, and nonlinear activations. With the convolutional layer, convolution operations are conducted to extract the features of input images. The convolutional layer with multiple kernels concurrently performs the convolutional operations to extract various feature images. The pooling layer following the convolutional layer is used to subsample and compress features, reduce the amount of calculation, and alleviate overfitting. Immediately afterwards, the fully connected layers are able to realize the full connection of parameters and generate the final classification results. Additionally, various nonlinear activation

functions are employed following the convolutional layers and the fully connected layers, aiming to lead the nonlinear properties to the network.

One convolution operation can be divided into two processes: (1) multiplication between elements in the kernel matrix and data matrix and (2) addition of all multiplication results, which can be expressed as follows:

$$y = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} x_{ij}, \quad (1)$$

where wij is the element of the convolution kernel, xij is the element of the input data, m is the row count, and n is the column count in the convolutional kernel. From Equation (1), one convolution operation is converted into vector-vector multiplication (VVM), and parallel convolution of multiple kernels is represented as a VMM [41]. At the same time, the optical matrix operation has been widely investigated [42], which makes it convenient to accelerate the convolution operation process with optical methods.

The optical architecture enabling the convolution operation has been blooming. For most of the reported optical CNNs, the convolution operations of CNNs are accelerated in the optical domain, and the rest remain in the electrical domain, which absorbs both the respective advantages for the ultra-bandwidth and low loss of light and the high precision recognition of electricity. Optical CNNs based on the implementation principle are generally divided into four categories: diffraction-based optical CNNs, interference-based optical CNNs, wavelength division multiplexing-based (WDM-based) optical CNNs, and tunable optical attenuation-based optical CNNs.

As an emerging research direction, optical CNNs have garnered significant attention since their inception. **Table 1** presents a comparison of four optical CNN schemes, highlighting their parallelism, computing speed, integration density, and reconfigurability.

Table 1. Comparison of optical CNN schemes.

Type	Parallelism	Computing Speed	Integration Density	Reconfigurability
Diffraction	**high**	**high**	**low**	**low**
Interference	**low**	**medium**	**medium**	**high**
WDM	**high**	**high**	**high**	**high**

Diffraction-based optical CNNs exhibit advantages in terms of parallelism, computing speed, and scalability due to their utilization of spatial light. The presence of spatial light allows for a large number of neurons in each layer, facilitating the expansion of multiple channels and kernels in the spatial domain, thus enabling high parallelism. The abundance of neurons and high parallelism contributes to achieving high computing speeds. However, diffraction-based optical CNNs also suffer from notable disadvantages. The discrete components used make the system bulky, and attempts at integration result in performance degradation. Moreover, kernel-loading devices such as DOE and metamaterials are nearly impossible to reconfigure, while SLM and DMD devices have low data rates (typically ~kHz).

An interference-based optical CNN excels in reconfigurability. This scheme often utilizes MZI for kernel matrix loading, enabling rapid refresh rates in the range of tens of GHz. Despite the advantages of high-speed reconstruction offered by MZIs, their relatively large volume limits the integration density of interference-based schemes. Furthermore, the use of coherent light restricts the transmission of only one light at a time in the optical waveguide, thereby constraining parallelism and computing speed.

WDM-based optical CNN represents a promising and extensively researched solution. This scheme fully exploits the optical wavelength dimension, leading to high parallelism. The use of MRR as a wavelength-sensitive optical attenuator, with its radius as small as several micrometers and modulation rate reaching tens of GHz, further enhances computing speed, integration density, and reconfigurability.

The optical CNN based on tunable optical attenuation depends on the specific characteristics of the optical attenuator used. Due to the absence of a unified conclusion, it is not listed in the table. Additionally, apart from the four types of optical CNNs discussed above, there are other noteworthy optical CNN solutions, such as photon frequency synthesis [43] and photodetectors with adjustable responsivity [44], which warrant further research.

Presently, although optical CNNs exhibit advantages in terms of bandwidth, latency, and computational speed compared with electrical architectures, optical CNNs face challenges in surpassing the limitations of small realizable matrix sizes, a limited range of realizable functions, and low computing precision. Consequently, extensive efforts are required for optical CNNs to gain widespread usage.

First, on-chip large-scale integration needs to be broken through. The reported on-chip integrated optical accelerated computing architectures only realize the integration of tens of thousands devices at present, which is far less than their electrical counterparts. In optical computing, the power required for calculations such as electro-optical conversion, photoelectric conversion, and analog-to-digital conversion remains basically unchanged. By integrating more photonic devices, the common power required for computing will be averaged, thereby improving the energy efficiency of optical computing and giving full play to the advantages of optical computing.

Second, more functions should be realized with optical methods. Most optical CNN solutions primarily focus on the optical implementation of convolutions. Although there are related studies on the optical implementation of nonlinear activation and full connection [35][38][45][46][47][48][49][50], these studies remain relatively limited and warrant further exploration. In particular, limited by the small matrix size, it is still a challenge to achieve large-scale full connections using optical methods. Implementing more functions using optical methods will be conducive to expanding the application field of optical computing and promoting the practical application of optical computing.

Finally, the development of an in situ trainable arbitrary reconfigurable computing architecture is essential. At present, most optical CNN implementations adopt the offline training method, and the weight matrix is pretrained in the neural network simulation model. As a result, a deviation between the simulation model and the experimental system inevitably appears. In situ training, which updates the weight directly and performs the computation at the original place, offers a new form to accelerate the reconfiguration performance of the neural network and improve its precision. Several recently proposed in situ training schemes, such as physics-aware training [51], adaptive training [52], and other relevant methods [53][54], have been successfully introduced to optical computing. These approaches can be effectively incorporated into the optical CNN framework. By incorporating in situ training, it is more convenient to reconstruct the network structure (such as changing the size and number of convolution kernels, changing the number of convolution layers, etc.), and the influence of errors can be considered during the training process. In this way, the refactoring of more different, more complex classification tasks on the same hardware will become a reality, rather than the simple tasks that are fixed today (such as classifying handwritten digits).

References

1. Fukushima, K.; Miyake, S. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. Compet. Coop. Neural Nets 1982, 36, 267–285.
2. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Comput. 1989, 1, 541–551.
3. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278–2324.
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
5. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the Computer Vision —ECCV 2014, Cham, Switzerland, 6–12 September 2014; pp. 818–833.
6. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2015, arXiv:1409.1556.
7. Szegedy, C.; Wei, L.; Yangqing, J.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
9. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. Science 2006, 313, 504–507.

10. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436–444.
11. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
12. Lawrence, S.; Giles, C.L.; Tsoi, A.C.; Back, A.D. Face recognition: A convolutional neural-network approach. IEEE Trans. Neural Netw. 1997, 8, 98–113.
13. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
14. Lee, H.; Pham, P.; Largman, Y.; Ng, A. Unsupervised feature learning for audio classification using convolutional deep belief networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 7–10 December 2009; pp. 1–9.
15. Abdel-Hamid, O.; Mohamed, A.R.; Jiang, H.; Penn, G. Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4277–4280.
16. Sainath, T.N.; Mohamed, A.; Kingsbury, B.; Ramabhadran, B. Deep convolutional neural networks for LVCSR. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 8614–8618.
17. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016, 529, 484–489.
18. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. Nature 2015, 518, 529–533.
19. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. Insights Imaging 2018, 9, 611–629.
20. Lakhani, P.; Sundaram, B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017, 284, 574–582.
21. Grigorescu, S.; Trasnea, B.; Cocias, T.; Macesanu, G. A survey of deep learning techniques for autonomous driving. J. Field Rob. 2020, 37, 362–386.
22. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review. IEEE Trans. Intell. Transp. Syst. 2022, 23, 722–739.
23. Data Is Giving Rise to a New Economy. Available online: https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy (accessed on 6 May 2017).
24. Kawatsu, C.; Koss, F.; Gillies, A.; Zhao, A.; Crossman, J.; Purman, B.; Stone, D.; Dahn, D. Gesture recognition for robotic control using deep learning. In Proceedings of the NDIA Ground Vehicle Systems Engineering and Technology Symposium, Detroit, MI, USA, 15–17 August 2017; pp. 1–7.
25. Schaller, R.R. Moore's law: Past, present and future. IEEE Spectr. 1997, 34, 52–59.
26. Moore, G.E. Cramming more components onto integrated circuits. Electronics 1965, 38, 82–85.
27. Theis, T.N.; Wong, H.S.P. The End of Moore's Law: A New Beginning for Information Technology. Comput. Sci. Eng. 2017, 19, 41–50.
28. Leiserson, C.E.; Thompson, N.C.; Emer, J.S.; Kuszmaul, B.C.; Lampson, B.W.; Sanchez, D.; Schardl, T.B. There's plenty of room at the Top: What will drive computer performance after Moore's law? Science 2020, 368, eaam9744.
29. Stone, H.S. A Logic-in-Memory Computer. IEEE Trans. Comput. 1970, 100, 73–78.
30. Patterson, D.; Anderson, T.; Cardwell, N.; Fromm, R.; Keeton, K.; Kozyrakis, C.; Thomas, R.; Yelick, K. Intelligent RAM (IRAM): Chips that remember and compute. In Proceedings of the IEEE International Solids-State Circuits Conference (ISSCC), San Francisco, CA, USA, 8 February 1997; pp. 224–225.
31. Sengupta, B.; Stemmler, M.B. Power Consumption During Neuronal Computation. Proc. IEEE 2014, 102, 738–750.
32. Miller, D.A.B. Attojoule Optoelectronics for Low-Energy Information Processing and Communications. J. Light. Technol. 2017, 35, 346–396.
33. Kitayama, K.-I.; Notomi, M.; Naruse, M.; Inoue, K.; Kawakami, S.; Uchida, A. Novel frontier of photonics for data processing—Photonic accelerator. APL Photonics 2019, 4, 090901.

34. Nahmias, M.A.; de Lima, T.F.; Tait, A.N.; Peng, H.T.; Shastri, B.J.; Prucnal, P.R. Photonic Multiply-Accumulate Operations for Neural Networks. IEEE J. Sel. Top. Quantum Electron. 2020, 26, 7701518.
35. Xu, X.; Tan, M.; Corcoran, B.; Wu, J.; Boes, A.; Nguyen, T.G.; Chu, S.T.; Little, B.E.; Hicks, D.G.; Morandotti, R.; et al. 11 TOPS photonic convolutional accelerator for optical neural networks. Nature 2021, 589, 44–51.
36. Bai, B.; Yang, Q.; Shu, H.; Chang, L.; Yang, F.; Shen, B.; Tao, Z.; Wang, J.; Xu, S.; Xie, W.; et al. Microcomb-based integrated photonic processing unit. Nat. Commun. 2023, 14, 66.
37. Meng, X.; Zhang, G.; Shi, N.; Li, G.; Azaña, J.; Capmany, J.; Yao, J.; Shen, Y.; Li, W.; Zhu, N.; et al. Compact optical convolution processing unit based on multimode interference. Nat. Commun. 2023, 14, 3000.
38. Sludds, A.; Bandyopadhyay, S.; Chen, Z.; Zhong, Z.; Cochrane, J.; Bernstein, L.; Bunandar, D.; Dixon, P.B.; Hamilton, S.A.; Streshinsky, M.; et al. Delocalized Photonic Deep Learning on the Internet's Edge. Science 2022, 378, 270–276.
39. Feldmann, J.; Youngblood, N.; Karpov, M.; Gehring, H.; Li, X.; Stappers, M.; Le Gallo, M.; Fu, X.; Lukashchuk, A.; Raja, A.S.; et al. Parallel convolutional processing using an integrated photonic tensor core. Nature 2021, 589, 52–58.
40. Li, X.; Zhang, G.; Huang, H.H.; Wang, Z.; Zheng, W. Performance Analysis of GPU-Based Convolutional Neural Networks. In Proceedings of the International Conference on Parallel Processing (ICPP), Philadelphia, PA, USA, 16–19 August 2016; pp. 67–76.
41. Vasudevan, A.; Anderson, A.; Gregg, D. Parallel Multi Channel convolution using General Matrix Multiplication. In Proceedings of the IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP), Seattle, WA, USA, 10–12 July 2017; pp. 19–24.
42. Abushagur, M.A.G.; Caulfield, H.J. 7—Optical Matrix Computations. In Optical Processing and Computing; Arsenault, H.H., Szoplik, T., Macukow, B., Eds.; Academic Press: Cambridge, MA, USA, 1989; pp. 223–249.
43. Fan, L.; Zhao, Z.; Wang, K.; Dutt, A.; Wang, J.; Buddhiraju, S.; Wojcik, C.C.; Fan, S. Multidimensional Convolution Operation with Synthetic Frequency Dimensions in Photonics. Phys. Rev. Appl. 2022, 18, 034088.
44. Zhen, W.; Zhou, X.; Weng, S.; Zhu, W.; Zhang, C. Ultrasensitive, Ultrafast, and Gate-Tunable Two-Dimensional Photodetectors in Ternary Rhombohedral ZnIn2S4 for Optical Neural Networks. ACS Appl. Mater. Interfaces 2022, 14, 12571–12582.
45. Yang, Z.; Tan, W.M.; Zhang, T.J.; Chen, C.D.; Wang, Z.X.; Mao, Y.; Ma, C.X.; Lin, Q.; Bi, W.J.; Yu, F.; et al. MXene-Based Broadband Ultrafast Nonlinear Activator for Optical Computing. Adv. Opt. Mater. 2022, 10, 2200714.
46. Ashtiani, F.; Geers, A.J.; Aflatouni, F. An on-chip photonic deep neural network for image classification. Nature 2022, 606, 501–506.
47. Williamson, I.A.D.; Hughes, T.W.; Minkov, M.; Bartlett, B.; Pai, S.; Fan, S. Reprogrammable Electro-optic Nonlinear Activation Functions for Optical Neural Networks. IEEE J. Sel. Top. Quantum Electron. 2020, 26, 1–12.
48. Zuo, Y.; Li, B.H.; Zhao, Y.J.; Jiang, Y.; Chen, Y.C.; Chen, P.; Jo, G.B.; Liu, J.W.; Du, S.W. All-Optical neural network with nonlinear activation functions. Optica 2019, 6, 1132–1137.
49. Guo, X.; Barrett, T.D.; Wang, Z.M.; Lvovsky, A.I. Backpropagation through nonlinear units for the all-optical training of neural networks. Photonics Res. 2021, 9, B71–B80.
50. Filipovich, M.J.; Guo, Z.; Al-Qadasi, M.; Marquez, B.A.; Morison, H.D.; Sorger, V.J.; Prucnal, P.R.; Shekhar, S.; Shastri, B.J. Silicon photonic architecture for training deep neural networks with direct feedback alignment. Optica 2022, 9, 1323–1332.
51. Wright, L.G.; Onodera, T.; Stein, M.M.; Wang, T.; Schachter, D.T.; Hu, Z.; McMahon, P.L. Deep physical neural networks trained with backpropagation. Nature 2022, 601, 549–555.
52. Zhou, T.; Lin, X.; Wu, J.; Chen, Y.; Xie, H.; Li, Y.; Fan, J.; Wu, H.; Fang, L.; Dai, Q. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. Nat. Photonics 2021, 15, 367–373.
53. Zhou, T.; Fang, L.; Yan, T.; Wu, J.; Li, Y.; Fan, J.; Wu, H.; Lin, X.; Dai, Q. In situ optical backpropagation training of diffractive optical neural networks. Photonics Res. 2020, 8, 940–953.
54. Hughes, T.W.; Minkov, M.; Shi, Y.; Fan, S.H. Training of photonic neural networks through in situ backpropagation and gradient measurement. Optica 2018, 5, 864–871.