Optical Multi-Imaging-Casting Accelerator for Fully Parallel Universal Convolution Computing

Changhe Zhou (chazhou@mail.shcnc.ac.cn)
Shanghai Institute of Optics and Fine Mechanics
Guoqing Ma
Shanghai Institute of Optics and Fine Mechanics
Rongwei Zhu
Shanghai Institute of Optics and Fine Mechanics
Junjie Yu
Shanghai Institute of Optics and Fine Mechanics

Article

Keywords: optical computing, optical convolution, universal convolution accelerator

Posted Date: October 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-870558/v1

Abstract

Over the last few years, optical computing has become a potential solution to computationally heavy convolution, aimed at accelerating various artificial intelligence applications. However, past schemes have never efficiently realized fully parallel optical convolution. Here, we propose a new paradigm for a universal convolution accelerator with truly massive parallelism and high precision based on optical multi-imaging-casting architecture. Specifically, a two-dimensional Dammann grating is adopted for the generation of multiple displaced images of the kernel, which is the core process for kernel sliding on the convolved matrix. Our experimental results indicate that the computing accuracy is typically close to 8-bit, and this accuracy can be improved further by using hybrid analog–digital coding method. In addition, a convolutional neural network for the standard MNIST dataset is demonstrated, and the recognition accuracy for inference is up to 97.3%. The paradigm reported here will open new opportunities for high-throughput universal convolution accelerators for real-time or quasi-real-time AI applications.

Introduction

A convolutional neural network (CNN), as “convolutional” implies, involves extensive convolution operations among neighboring layers, followed by batch normalization and nonlinear activation for expected performance[1–3]. Interestingly, these massive linear multiply–accumulate (MAC) operations account for more than 80% of the total number of calculations for deep neural networks (DNNs)[4]. Thus, the convolution operation, which is not suitable for modern advanced electric serial processors, is becoming the heaviest burden for artificial intelligence (AI) algorithms. In addition, with an increase in the scale of the network, the computational overhead of convolution operations also increases exponentially. It has been shown that the amount of computing power required to train state-of-the-art DNNs has doubled every 3.5 months[5], which is far beyond that of traditional electrical integrated circuits (EICs) following Moore’s law. Although the calculation of convolution can be accelerated by using parallel electrical coprocessors, such as graphics processing units (GPUs) and tensor processing units (TPUs), it is still difficult to handle millions of MAC operations in a fully parallel manner for DNNs practically[6,7]. In contrast, it has been proven that a large number of MAC operations can be simultaneously executed during a single pass of light, and this may be the prime motivation for the recent interest in optical computing[8,9]. In fact, photonic solutions for computing have been investigated for at least 70 years[10–12]. However, compared with fast-growing EICs, the development of optical computing gradually slowed down in the late 2000s[10] because of the lack of application-driven motivation and adequate optical convolution architectures.

Recently, due to the remarkable achievements in AI, there has been a revived interest in trying to improve computing power, energy efficiency, and processing speed by exploiting photonic or hybrid optical–electric processors instead of their electronic counterparts[8,9,13–15]. Two mainstream architectures for optical computing have been rapidly developed. One is based on a planar waveguide on a two-dimensional (2D) substrate[16,17], and the other is realized by multiple cascading diffractive optical elements (DOEs) in three-dimensional (3D) space[18]. However, the planar architecture, including both Mach–Zehnder interferometers[16] and microring resonators[19,20] does not make full use of the 3D interconnection capability of optics, whereas the 3D architecture requires full manipulation of the electromagnetic field with high precision; thus, it will be challenging to fabricate high-precision subwavelength DOEs in 3D space[21].

Although it was predicted that photonic processors have the potential to be at least ten thousand times faster than state-of-the-art EICs[13,14], the past schemes have not realized fully parallel convolutional computing compared with their electronic counterparts, especially when high precision is required. Here, we propose a new paradigm for a universal convolution accelerator with full parallelism and adequate precision based on optical multi-imaging-casting (OMica) architecture, which is capable of calculating the arbitrary-encoded hybrid analog–digital matrix convolution. The architecture can be seen as the starting point of a new roadmap for optical computing, with potential for building fully massively parallelized optical convolution accelerators to overcome the intrinsic shortage of computing power and unsatisfactory energy efficiency of

EICs. Furthermore, it promises to give full play to the advantages of AI algorithms or accelerate other practical applications where rapid big data processing is desired.

Results

OMica Accelerator Architecture

A schematic of the OMica architecture proposed here for the acceleration of the universal matrix convolution is shown in Fig. 1. The light that carries the information of kernel *A* is replicated into multiple beams with different angular orientations by a 2D beam splitter, and thus, each sub-beam carries all the information of kernel *A* and propagates along its own diffraction angle. Next, these sub-beams pass through a 4F lens system, which comprises a pair of Fourier lenses, L_1 and L_2, and projects multiple displaced images of kernel *A* onto its conjugate plane at a magnification of ´1, where the feature map, matrix *B*, is located. Thus, once the sub-beams carrying information of kernel *A* pass through matrix *B*, element-wise multiplication is simultaneously realized. The kernel's sliding process is automatically executed with massive parallelism if the appropriate distance *d* between matrix *A* and the beam splitter is selected, making the displaced distance equal to the element spacing of matrix *B* (see Supplementary Note 1). Finally, each sub-beam carrying the information of the element-wise multiplication of displaced kernels *A* and matrix *B* are truncated by the aperture of the modulator located in matrix *B*, and the passed field is focused by another convergent lens L_3 to execute the accumulation operation. Moreover, the accumulations for these sub-beams are inherently separated in the Fourier plane of lens L_3 due to the difference in the angular spectra for these sub-beams; thus, the final convolution matrix *C* is obtained from the 2D spot array on the Fourier plane, that is,

$$C = A \odot B \quad (1)$$

where '⊙' means convolution operation. In addition, owing to the object–image conjugate configuration, the OMica accelerator proposed here possesses a large space-bandwidth product[22-2433], which makes it possible to realize massive parallelism with sufficiently high accuracy.

In our proof-of-concept implementation, a homemade 2D even Dammann grating (DG) was inserted into a 4F system (Fig. 1(a)), working as a beam splitter for the generation of multiple displaced images of convolution kernel *A* (see Methods and Supplementary Notes 1 & 5). Here, the even DG is the critical element for the realization of simultaneous translational sliding of kernel *A* on feature map *B*. Two spatial light modulators (SLMs) are located on the object and image planes of the 4F system, where the two convolution matrices, kernel *A* and feature map *B*, are dynamically loaded. In the experiment, the light intensity was used as the information carrier, and two amplitude-only SLMs were used to embed the information of Kernel *A* and feature map *B* into the incident uniform light beam. Therefore, in principle, only nonnegative matrices can be loaded and calculated based on this hardware. To overcome this drawback, an arbitrary digital encoded method was developed for hybrid analog–digital optical convolution computing. In a hybrid analog–digital framework, one can easily decompose a negative arbitrary-bit matrix, in the same form as a positive matrix, into one larger-scale or several same-size low-bit matrices in spatial or temporal sequences, respectively[25,26]. In other words, both positive and negative numbers in the original matrix can be expressed as $a = \sum_{n=0}^{N-1} c_n(-2)^n$, where $\{c_n\}$ are *N*-separated *k*-bit bytes, and each $\{c_n\}$ denotes a *k*-bit number, with $n = 0, 1, 2, \ldots, N-1$. After this decomposition, a negative arbitrary-bit matrix is transformed into low-bit non-negative matrices, and it is possible to load these matrices on the SLMs. The principle of this encoding method is shown schematically in Fig. 1(b). Notably, there is a balance between computing precision and computing power, which can be tuned by changing the parameter *k*. A small *k* suggests that a higher precision and a lower computing power will be generated, whereas a large *k* indicates a large computing power and relatively low precision. Therefore, compared with analog optical convolutional computing, this encoding method can improve the computing precision to the same extent[26].

Here, as an example, the encoding process is demonstrated step-by-step for a matrix with elements having a quaternary (2-bit) number under the condition of $k = 1$. First, the quaternary number for each element of the high-bit matrix to be encoded is expressed in multiple low-bit elements after encoding. For example, the first element is written as $-2 = 0 \times (-2)^2 + 1 \times (-2)^1 + 0 \times (-2)^0$. Thus, each element of the matrix to be encoded is expressed as multiple elements in the encoded matrix. Therefore, the elements of the matrix are arranged in rows after encoding, denoted as P_1, P_2, P_3, P_4, P_5, and each element in the column direction is encoded with three bytes, denoted as Bit_3, Bit_2, and Bit_1, as shown in Fig. 1 (b). For example, the first element, −2, is expressed as {010} in the first column of the encoded matrix, that is, $c_2 = 0$, $c_1 = 1$, and $c_0 = 0$. Then, the converted matrices are sequentially loaded onto the SLMs for computing in a temporal or spatial sequence. Notably, in a spatial sequence, some zero elements should be inserted into the encoded matrix between two adjacent rows or columns of the original high-bit matrix to avoid aliasing. In this situation, the physical pixels of the SLMs will not be fully utilized because of the redundant zero elements. On the other hand, the encoding method in temporal sequence takes full advantage of the physical pixels of SLMs. However, the convolution must be executed between all bits of either kernel *A* or matrix *B*, and thus the refresh rate of the system will be at a large discount. Therefore, a compromise should be struck between high computing power and high computing precision by choosing an appropriate parameter *k* when OMica hardware is used for computing acceleration in the hybrid analog–digital framework described above.

Hybrid Analog–Digital Coding Matrix Convolution

As an example, the hybrid analog–digital optical convolution of 10 pairs of random binary matrices, 10 pairs of quaternary matrices, and 10 pairs of quaternary matrices with negative elements are demonstrated. In our proof-of-concept experiment, the maximum matrix size loaded into OMica hardware is about 10 × 10 due to the finite signal-to-noise ratio caused by the finite contrast of reflective liquid crystal SLMs available. Fig. 2 compares the experimental results of the optical convolution of a pair of binary matrices and two pairs of quaternary matrices with the theoretical results. The results, typical of these three types of matrix convolutions, are shown in Figs. 2(a) – (c). In each box, the theoretical results obtained by an electric computer (full precision, 64-bit) are illuminated in the first subfigure of the first row. The light intensity distributions of the spot arrays on the detected plane, denoting the raw results of the convolution, are shown in the second subfigure, and the experimental results before decoding are shown in the third subfigure. The absolute error map, defined as $|\boldsymbol{C}_{theo} - \boldsymbol{C}_{exp}|$, is shown in the last subfigure of the first row, where $\boldsymbol{C}_{theo}$ and $\boldsymbol{C}_{exp}$ are the theoretical and experimental convolutional results, respectively, and "|.|" denotes the absolute operation. In addition, the theoretical and experimental results of the convolution after decoding are shown in the first and second subfigures in the second row. It is shown that the overall trend of the experimental and theoretical results of the convolution is consistent.

Fig. 2(a) shows the results of the convolution of two 10 × 10 binary matrices. It can be seen that the mean value of the absolute errors $|\boldsymbol{C}_{theo} - \boldsymbol{C}_{exp}|$ is 0.240, and the maximum error is below 0.4, indicating that the computing accuracy after digitalization is 100%. Fig. 2(b) shows the results of the convolution of two 3 × 10 2-bit matrices. The mean value of the absolute errors is 0.114, and it is seen that the maximum value is approximately 0.239 before decoding, which indicates that high precision is achieved by the OMica architecture. Fig. 2(c) shows the results of the convolution of two 2 × 10 2-bit matrices with negative elements. The mean error is 0.080, and the maximum value is approximately 0.145. It should be noted that the mean error before decoding is greater than that of the other two cases, mainly due to the increased crosstalk resulting from relatively large convolution elements. Moreover, the other two encoded matrices are filled with zero elements to avoid aliasing, which further reduces the crosstalk and final error. Because the maximum absolute errors are all less than 0.5 for these three cases, the correct convolution results, with an accuracy of 100%, can still be obtained after digitalization. Thus, the experimental light intensity distribution of the three cases precisely reflects the values of the convolution results.

The error distribution of all 30 sets of matrix convolutions is shown in Fig. 2(d), and it is clear that the maximum absolute error is less than 0.5. This means that no errors will occur after the convolution results are digitalized, suggesting good reliability and robustness of the OMica architecture. Notably, the accuracy is related to the stability of the light source,

contrast of the modulators, transferring ability of the imaging system, and sensitivity and dynamic range of the detector. In addition, we experimentally demonstrated the convolution results of larger-scale and higher-bit matrices, as shown in Figs. S12, S13, S14, and S15 (see Supplementary Note 6), where the convolution results of 3 × 3 1-bit and 20 × 20 1-bit matrices, two 10 × 10 8-bit matrices, 20 × 20 8-bit matrices, and 180 × 224 8-bit matrices are given, respectively.

CNNs based on MNIST

Based on the abovementioned hybrid analog–digital coding method, we demonstrate the recognition of handwritten digits based on the OMica architecture. Here, a binary neural network (BNN)[27] is implemented as an example to test the robustness and accuracy of the proposed optical hardware. For a BNN, the input signal is a binary (0 or 1) image, and the kernel is a binary matrix with a weight of −1 or +1[28]. Each kernel of the BNN trained in advance is divided into two sub-matrices; one is a low-bit (positive) matrix and the other is a high-bit (negative) matrix, as shown in Fig. 3(a). Intuitively, it seems that two convolution operations should be executed in the temporal sequence. Interestingly, 10 original kernels need to be divided into 10 low-bit sub-kernels and another uniform high-bit sub-kernel. Furthermore, the first positive kernel and the negative kernel are exactly the same; thus, the total number of convolution kernels after encoding is still 10, which means that no additional computational overhead is incurred. The final convolution result can be obtained by the addition of the positive and negative convolution result multiplied by −2. Fig. 3(b) shows the inference process of the CNN based on encoding low- and high-bit kernels. The 10 encoded kernels are sequentially loaded onto the SLM located at the input plane of matrix *A*, and the binary input images with a scale of 28 × 28 are loaded sequentially onto the SLM located at the input plane of matrix *B*. When light passes through the two SLMs in sequence, and is then focused and separated by the focusing lens, the spot array denoting the convolution results is captured by the detector on the focal plane. Finally, the original convolution results are obtained by decoding the corresponding low- and high-bit convolutions.

Fig. 3(c) shows the error map between the theoretical and experimental results of an input image of a handwritten digit 7 convolved by the first kernel. Compared with the above three examples with an input matrix of 10 × 10, the size of a standard input image of handwritten digits is 28 × 28, whereas the size of the convolution kernel is almost the same, and the average value of the absolute errors is 0.405. This suggests that it is possible to calculate the optical convolution of larger-scale matrices using the OMica architecture with high precision. The following pooling layer, nonlinear operations, and full connections are executed by a classical electrical computer.

To validate the reliability and robustness of the system, we implemented blind-testing for the first 1000 sets of MNIST images with serial numbers from 1 to 1000. The experimental results indicate that a blind-testing accuracy of up to 97.3% was achieved for the OMica convolution accelerator, whereas the recognition accuracy was only 96.7% for the same test dataset for electrical computers. This was due to the computing error of the optical convolution also carrying characteristics of the input images, thus further strengthening the feature extraction ability. Noticeably, the error maps for different handwritten digits are highly correlated with the input image, as shown in Fig. 4. Therefore, the recognition accuracy of the optical convolution system based on the MNIST dataset was slightly higher than that of the electronic computer, as shown in Fig. 3(d). By further optimizing the kernel weights of the optical convolution system, direct training of the optical CNN is expected to yield better results than those of an electronic computer. On this basis, the architecture can be effectively used as a hardware accelerator with large computing power in various DNNs.

To the best of our knowledge, the OMica architecture is the only optical parallel acceleration solution capable of achieving both high-precision convolutional computers and AI hardware accelerators with high recognition accuracy. In addition, not only convolution layers but also the pooling layers and fully connected layers (all layers are linear convolution calculations) could be realized by the OMica architecture if an appropriate distance *d* (Fig. 1(a)) is chosen. For AI algorithms, it has been shown that very high accuracy is not required[29], especially for inference tasks. Inference models work nearly as well with 4–8 bits of precision and trained with nearly 8–16 bits of precision per computation[30]. Our results suggest that the computing precision is close to 8 bit; thus, it is sufficiently accurate for most AI inference applications. Moreover, the computing

accuracy could be improved by more than 8 bits if high-contrast modulators, such as DMDs, are employed, and this accuracy can be improved further by adopting hybrid analog-digital encoding method. Thus, the results obtained by this optical accelerator would be adequate for training most AI models. In addition, when training the neural network directly in the OMica system, the physical characteristics of the system itself are also trained, such as alignment errors and crosstalk, which are expected to further improve the performance of the neural network mentioned above.

An OMica accelerator for fully parallel universal convolution computing was proposed, and a hybrid analog–digital encoding scheme with sufficiently high precision was demonstrated. In principle, the convolution of an arbitrary bit matrix with massive parallelism and sufficient accuracy can efficiently be calculated by using a suitable encoding scheme and the OMica architecture. Moreover, the convolution is universal, and the computing results obtained may be easy to transplant to any other computing platform. Our proof-of-concept experimental results prove the feasibility of the optical convolution of 10 × 10 matrices with an accuracy above 8-bit, meaning that the results obtained by this optical accelerator are sufficiently accurate for most AI inference tasks, even for training some AI models. Furthermore, a BNN for recognition tasks of handwritten digits for the standard MNIST dataset was constructed, and the inference process was demonstrated based on this optical hardware. The results indicate that the blind-testing recognition accuracy is as high as 97.3%, which is even higher than that predicted by pure electrical networks. These proof-of-concept experimental results suggest that the OMica architecture can be used for massive parallelism, high-precision, and high-efficiency AI accelerators, and this computing paradigm has potential applicability in the construction of task-specific cloud computing centers or other AI computing centers. By developing high-speed SLMs with higher contrast, optimizing a special-proposed projection imaging system, and configuring a dedicated dot array lighting source, it is possible to construct a photonic coprocessor with higher computing power and lower energy consumption than state-of-the-art supercomputers, such as Fugaku, based on the proposed OMica architecture. In addition, the characteristics of the imaging system itself imply that the computing power of the system can be further increased by cascading multiple 4F systems and employing extra multiplexing degrees of freedom[31,32]; thus, a hybrid optical–electrical computer center or data center can be directly constructed. In the future, with the advancement of nonlinear optical elements[33-35], a scheme based on the OMica architecture could also be integrated into pure photonic accelerators by combining planar waveguides[36,37], metasurfaces[38-40], or some other technologies[41,42].

In summary, the OMica architecture is expected to be used in self-driving vehicles[43], machine version[44], and other fields that require large computing power for real-time or quasi-real-time data processing. This opens the door for increasing the computing power and energy efficiency for convolution by using high-performance devices, such as larger-scale modulators with higher updating frequencies and detectors or detector arrays with wider dynamic range and higher sampling frequencies, which, in the near future, would be superior to the most powerful supercomputers.

Discussion

As shown in Fig. 1(a), the planes of matrices *A* and *B*, confocal plane of the 4F system, and plane of the detector are in a conjugated object–image relationship with each other. When a beam splitter, such as a DG, is placed behind the plane of matrix *A*, the two pairs of object–image relationships mentioned above still hold. Each diffraction order of the beam splitter involved in the convolution is still imaged to the plane of matrix *B* when we adjust the suitable distance d_0 between matrix *A* and the beam splitter to match the diffractive angle of the beam splitter. Therefore, it is possible to greatly reduce the physical size of the matrix elements, which indicates that a much larger scale of the matrix, such as 1k × 1k, could be calculated in parallel. Under these conditions, the peak computing power of the OMica architecture will reach 10 peta (10^{15}) operations per second (POPS), which is even faster than a top GPU, such as the TITAN RTX (Nvidia)[45], if a modulator with a higher refresh rate (typically 10 kHz) is used, such as a DMD or a specially designed micro-electro-mechanical system (MEMS). Further, if other multiplexing methods, including polarization, wavelength, and spatial mode, are used, then speeds of at least 10–10^2 times faster than this estimation can be achieved[20,30,31]. Therefore, based on the OMica architecture, the computing power for convolution may be superior, or at least comparable, to that of the most powerful supercomputer[46]

(the peak performance of the top system, Fugaku, is over 10^3 POPS for AI applications) in the near future with larger scale and higher updating frequency devices.

In addition, the power consumption of the optical convolution computing system is much lower than that of an electronic processor with the same computing power, even for such a bulk optical system. This takes into account the operating power consumption of the optoelectronic device itself and assumes that the total power consumption of the entire OMica system, including the light source, two modulators and the detector, is less than 100 W. Moreover, if a more sensitive detection device, such as a multiphoton counter, is employed, the power consumption will decrease drastically[47]. In contrast, a powerful supercomputer is energy hungry, and its power consumption is typically as high as 10^4–10^5 kW (Fugaku's power is 28,355 kW). Evidently, the OMica architecture will consume far less power than supercomputers, whereas its computing power for a specific task (convolution) could be at least comparable to that of Fugaku, the top supercomputer.

Compared with the most popular scheme involving planar waveguides on a 2D substrate, noticeably, the OMica architecture inherently takes full advantage of the 3D connection ability of optics and thus, can potentially achieve higher computing power. Recently, Mario et al.[48] proposed an optical system that performs fast updating of optical neural networks based on two amplitude-only DMDs, where one amplitude-only DMD is located at the Fourier transform plane of the other. Although the mapping relationship between the input images and the recognition digits can be successfully established using this method, the computing results are essentially not standard convolutions; thus, this method cannot be used for high-precision universal convolution computing. Moreover, it is difficult to align the two DMDs pixel-by-pixel, and thus, it will be challenging to realize large-scale optical networks due to the inverse Fourier transform relationship between the input and the filter planes. Recently, Lin et al.[49] demonstrated a reconfigurable scheme for realizing a 3D architecture comprising multiple cascading DOEs, where two programmable modulators and a DMD and another pure-phase SLM were employed for amplitude and phase modulation, respectively. Because of the coherent working mode, micron-sized pixels, the alignment error between the DMD and SLM, and the alignment errors between different layers make it difficult to achieve high computing precision. Therefore, recognition is drastically degraded without adaptive training. Although this scheme performs well after adaptive training, it cannot be used for universal convolution computing due to its low precision.

In contrast, a $CMOS_2$ monitor camera can be added on the conjugating plane of two SLMs owing to the object–image conjugate relationship, and thus two SLMs can easily be aligned using a monitor camera. Moreover, an incoherent light source could be used in the OMica architecture, so that the sensitivity and speckle noise could be avoided. Therefore, the convolution accelerator enabled by the OMica architecture can be used for computing universal matrix convolution, and the results obtained by this hybrid optical–electrical hardware can be easily transferred to any other computing platform, including photonic, hybrid optical–electrical, and traditional electric processors or coprocessors. Owing to its universality, this architecture can be used for building task-specific cloud computing centers, or some other AI accelerating centers, even for this bulk optical system at present.

At present, only one kernel A and one input feature map B are loaded onto these two SLMs. It is also possible to load multiple kernels on the first SLM, and thus the convolutions between multiple kernels and multiple input channel feature maps could be realized in parallel by filling an appropriate number of zero elements between any two adjacent kernels. Furthermore, it is worth noting that considering the actual hardware scale, it is often necessary to split and reorganize the input feature map to further improve the hardware utilization, that is, to load different matrix combinations to the SLMs to execute the convolution process[50].

The current CMOS technology, in principle, is sufficient for developing high-quality devices, such as SLMs and detectors, for optical computing, although these task-specific devices are not currently available. This work provides a promising method of building optical convolution computing processors to overcome the intrinsic shortage of computing power and unsatisfactory energy efficiency in traditional electrical processors, and the experimental results verify the advantages of optical convolution systems for various application scenarios.

Materials And Methods

Experimental system. A schematic of the entire proof-of-concept experimental system is presented in Supplementary Fig. S2. The incoherent light beam is emitted from an LED chip (M450D3, Thorlabs) operating at a center wavelength of λ = 450 nm, passes through a lens group composed of L_1 (f = 300 mm) and L_2 (f = 300 mm), and projects its image onto the plane of a pinhole (AP_1) for filtering. Next, the passed beam is reflected by an aluminum mirror (M_1) and passes through lens L_3 (f = 300 mm) for expansion. The expanded beam is then divided into two parts using a cube polarization beam splitter (PBS_1). One part, with *s*-polarization, is directed onto a homemade aperture array (APA, pitch: 432 μm, aperture diameter: 360 μm), and then the transmitted field passes through a pair of confocal lenses, L_4 (f = 300 mm) and L_5 (f = 300 mm). The image of the APA is projected onto the first amplitude-only SLM_1 (the B plane, pixels: 8 μm, 1920 · 1080, Xi'an CAS Microstar), where the kernel matrix A is loaded. The modulated field reflected from SLM_1 is then transmitted through another 4F system, composed of a pair of Fourier lenses, L_6 (f = 300 mm) and L_7 (f = 300 mm). On the conjugating plane of SLM_1 (the C plane), another amplitude SLM_2 (pixels: 8 μm, 1920 · 1080, Xi'an CAS Microstar) is located. Critically, a homemade 20 · 28 DG (period: 225 μm) is inserted after the SLM_1 plane at a certain distance d_0, which is matched with the diffractive angle of each diffractive order of the DG. When the modulated field passes through the DG and is separated into sub-beams, each sub-beam carries all the modulated light information of matrix A, and all are imaged into the SLM_2 plane through the Fourier lenses group. Here, pinholes (AP_2 and AP_3) are inserted to filter out the high-frequency components of the light on the frequency-domain plane between the two 4F systems. Then, the light-carrying element-wise multiplication information of matrices A and B is reflected by SLM_2, and after being reflected by PBS_3, it passes through the lens group, L_8 (f = 150 mm) and L_9 (f = 300 mm). The image of the field on plane C is projected onto the plane of a square aperture (17.28 · 17.28 mm^2. After passing through the square aperture, the light corresponding to the zero-filling part of matrix B is truncated, and a spot array denoting the convolution is detected by a scientific CMOS camera ($sCMOS_1$, pixels: 11 μm, 2048 · 2048, Dhyana 95, Xintu Optoelectronics) near the focal plane of lens L_{10} (f = 300 mm). After postdata processing, the convolution matrix data are obtained by the accumulation of the intensity signal for each spot. Here, an optical attenuator and a narrowband filter (at 450 nm) were mounted before the $sCMOS_1$ camera to improve the signal-to-noise ratio. To align two SLMs pixel-by-pixel, another cube non-polarization beam splitter (BS) is inserted before the $sCMOS_1$ camera and another lens L_{12}, followed by another camera ($CMOS_2$, pixels: 4.8 μm, 1920×1200, EO2323), which is used for monitoring. By adjusting the location of lens L_{12}, the images of matrices A and B loaded onto these two SLMs can be clearly projected onto the $CMOS_2$ camera, and the alignment error between these two SLMs can be seen explicitly through this monitor camera. Moreover, when adding a pinhole on the confocal plane of lenses L_6 and L_7 to allow only one diffraction order of DG to pass through sequentially, we can directly observe the complete convolution sliding process on the $CMOS_2$ (see Supplementary Note 2 and Fig. S5). Another part of the beam is reflected by aluminum mirrors (M_2, M_3, M_4, and M_5) and then directed into lens L_{11}. The focused field is recombined by the BS located before the $sCMOS_1$ camera and then impinged onto the detection plane of the camera. This optical signal is used to compensate for the temporal intensity fluctuation of the LED light source.

CNN model architecture and training. The CNN adopted here is a BNN, where the input signal is a binary (0 or 1) image, and the kernel is also a binary matrix with a weight of − 1 or + 1. The BNN contains a total of five layers: an input layer, a convolutional layer, pooling layer, fully-connected layer, and an output layer, as shown in Supplementary Fig. S7. The input layer consists of binary images of 28 × 28 handwritten digits. Unlike an electronic computer that needs to rearrange the input grayscale image into a one-dimensional vector, the optical convolution computing system directly loads the input 2D image onto SLM_1, which gives full play to the numerous advantages of optical spatial interconnection. Next, 10 9 × 9 convolution kernels are used to execute the convolution operation on the input image. ReLU is chosen for the nonlinear activation function after convolution. The size of the feature maps after convolution and the nonlinear operation is 20 × 20. The average pooling layer is then used to further extract the feature information, and the size of the feature maps is 10 × 10. Next, the pooled feature map is flattened into a 1 × 100 vector, followed by two fully-connected layers with 200 and 10 neurons. Finally, the ReLU and sigmoid nonlinear activation operations follow after the above two fully connected layers,

respectively. The final output layer is the proportion of input digits in 10 categories, and the highest proportion is the classification result. The learning rate of the network was set to 0.01, and the training batch size was set to 50. The parameters of the BNN were initialized by setting the initial value of the weight to − 1 or + 1. Before the training procedure, the 60,000 image samples were shuffled to generate a reasonable gradient for accelerating the network convergence, and the 60,000 image samples were divided into 80% as the training set and 20% as the validation set. The number of variables in the BNN was 3.5×10^5. During the training procedure, the weights were constrained between binary values of − 1 or + 1. The number of training epochs was 4. The training set and the validation set highly overlapped in the training curve, which proves that the model has good learning and generalization performance (Supplementary Fig. S8). The simulation results on an electrical computer suggest that, for 10,000 test samples from the MNIST dataset, the recognition accuracy of blind-testing on the first 1,000 test samples (with serial numbers from 1 to 1000) was 96.7%, and it was 96.3% for all 10,000 test samples.

Data processing. The entire data postprocessing flowchart can be described as follows. First, median filtering is performed on the intensity map detected by the $sCOMS_1$ camera, and the image tilt is corrected to simplify handling the matrices. Second, the area of interest (AOI) containing the target convolution matrix information is tailored, and the background noise is subtracted to further improve the signal-to-noise ratio. Third, the center of each spot is determined by calculating the centroid coordinates in the spot array image. Subsequently, the gray scale of all pixels within a circle is summed; the centroid is the center of the circle, and its radius is chosen to minimize the calculation errors. Fifth, by normalizing the intensity of the light spot divided by the calibrated light intensity and mapping the intensity map to the theoretical convolution results yields the experimental matrix values for convolution. Next, the absolute error map is obtained by calculating the absolute value of the difference between the experimental and theoretical convolution results before decoding. Finally, the convolution results are decoded into a high-bit matrix, exhibiting the correct grayscale map, if necessary. Similarly, the error map after decoding is calculated.

Declarations

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper. Additional data used in this study are available from the corresponding author upon reasonable request.

Supplementary information accompanies the manuscript on the Light: Science & Applications website (http://www.nature.com/lsa).

Acknowledgment

This work was supported by the Cutting-Edge Sciences Important Research Program, Bureau of Frontier Sciences and Education, Chinese Academy of Sciences under Grant QYZDJ-SSW-JSC014; and in part by the Shanghai Science and Technology Committee under Grants 19DZ2291102, 19JC1415400, and 20ZR1464700. The authors appreciate the critical discussion on this concept and also the experiment with Guowei Li.

Author information

Contributions

C.H.Z conceived the idea. J.J.Y designed the experiments. G.Q.M and R.W.Z performed the experiments. G.Q.M and J.J.Y analyzed and simulated the data. C.H.Z contributed materials/analysis tools and supervised the project. G.Q.M and J.J.Y co-wrote the paper.

Corresponding authors

Correspondence to Junjie Yu or Changhe Zhou

Correspondence and requests for materials should be addressed to *Junjiey@siom.ac.cn, chazhou@mail.shcnc.ac.cn*

Ethics declarations

Conflict of interest

C.H.Z, J.J.Y, and G.Q.M claim a Chinese patent on the architecture and devices presented in this work through the Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences at Shanghai.

References

1. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems* **25**, 1097–1105 (Curran Associates, 2013).
3. LeCun, Y., Bengio, Y. & Hinton. G. Deep learning. *Nature* **521**, 436–444 (2015).
4. Li, X. et al. Performance analysis of GPU-based convolutional neural networks. In *45th International Conference on Parallel Processing* (ICPP) 67–76 (IEEE, 2016).
5. De Lima, T. F. et al. Machine Learning With Neuromorphic Photonics. *J. Lightwave Technol.* **37**, 1515–1534 (2019).
6. Ito, Y., Matsumiya, R. & Endo, T. ooc_cuDNN: Accommodating convolutional neural networks over GPU memory capacity. *IEEE International Conference on Big Data (Big Data)* 183–192 (2017).
7. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vision Pattern Recognit (CVPR)* 770–778 (2016).
8. Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47, (2020).
9. Shastri, B. J. et al. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photon.* **15**, 102–114 (2021).
10. Ambs, P. Optical Computing: A 60-Year Adventure. *Adv. Opt. Photonics* 2010, 1–15 (2010).
11. Weaver, C. S. & Goodman J. W. A technique for optically convolving two functions. *Appl. Opt.* **5**, 1248–1249 (1966).
12. Jutamulia, S. & Yu, F.T.S. Overview of hybrid optical neural networks. *Opt. Laser Tech.* **28**, 59–72 (1996).
13. Prucnal, P. R. & Shastri, B. J. Neuromorphic Photonics. (Boca Raton: CRC Press, USA, 2017).
14. Thomas, F., Shastri, B. J., Tait, A. N., Nahmias, M. A. & Prucnal, P. R. Progress in neuromorphic photonics. *Nanophotonics* **6**, 577–599 (2017).
15. Zhang, Q., Yu, H., Barbiero, M., Wang, B. & Gu, M. Artificial neural networks enabled by nanophotonics. *Light Sci. Appl.* **8**, 42–56 (2019).
16. Shen, Y. C. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
17. Prabhu, M. et al. Accelerating recurrent Ising machines in photonic integrated circuits. *Optica* **7**, 551–558 (2020).

18. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
19. Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H. & Pernice, W. H. P. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
20. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
21. Soukoulis, C. M. & Wegener, M. Past Achievements and Future Challenges in the development of 3D photonic metamaterials. *Nat. Photon.* **5**, 523–530 (2011).
22. Ozaktas, H. M. & Urey, H. Space-bandwidth product of conventional Fourier transforming systems. *Opt. Commun.* **104**, 29–31 (1993).
23. Greenbaum, A. et al. Increased space-bandwidth product in pixel super-resolved lensfree on-chip microscopy. *Sci. Rep.* **3**, 1717–1725 (2013).
24. Lohmann, A. W. et al. Space-bandwidth product of optical signals and systems. *J. Opt. Soc. Am.* **13** 470–473 (1996).
25. Zhou, C. H., Liu, L. R. & Wang Z. J. Binary-encoded vector–matrix multiplication architecture. *Opt. Lett.* **17**, 1800–1802 (1992).
26. Liu, L. R., Li, G. Q. & Yin, Y. Z. Optical complex matrix–vector multiplication with negative binary inner products. *Opt. Lett.* **19**, 1759–1761 (1994).
27. Qin, H. T. et al. Binary neural networks: A survey. *Pattern. Recogn.* **105**, 107281 (2020).
28. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R. & Bengio, Y. Binarized Neural Networks Training Neural Networks with Weights and Activations Constrained to + 1 or − 1. *arXiv*:1602.02830 (2016).
29. Gupta, S., Agrawal, A., Gopalakrishnan, K. & Narayanan, P. Deep Learning with Limited Numerical Precision. *arXiv*:1502.02551 (2015).
30. Nahmias M. A. et al. Photonic Multiply-Accumulate Operations for Neural Networks. *IEEE J Quantum Electron* **26** (1), 7701518–7701536 (2020).
31. Xu, X. Y. et al. 11 TOPS photonic convolutional acceleraor for optical neural networks. *Nature* **589**, 44–51 (2021).
32. Wang, J. et al. Terabit free-space data transmission employing orbital angular momentum multiplexing. *Nat. Photon.* **6**, 488–496 (2012).
33. Jha, A., Huang, C. & Prucnal, P. R. Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics. *Opt. Lett.* **45**, 4819–4822 (2020).
34. Zuo, Y. et al. All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132–1137 (2019).
35. Ryou, A. et al. Free-space optical neural network based on thermal atomic nonlinearity. *Photon. Res.* **9**, 128–134 (2021).
36. Gruber, M. Multichip module with planar-integrated free-space optical vector-matrix-type interconnects. *Appl. Opt.* **43**, 463–470 (2004).
37. Mínguez-Vega, G., Gruber, M., Jahns, J. & Lancis, J. Achromatic optical Fourier transformer with planar-integrated free-space optics. *Appl. Opt.* **44**, 229–235 (2005).
38. Zhang, Y. et al. Electrically reconfigurable non-volatile metasurface using low-loss optical phase-change material. *Nat. Nanotechnol.* **16**, 661–666 (2021).
39. Wu, Z., Zhou, M., Khoram, E., Liu, B. & Yu, Z. Neuromorphic metasurface. *Photon. Res.* **8**, 46–50 (2020).
40. Kwon, H., Sounas, D., Cordaro, A., Polman, A. & Alù, A. Nonlocal metasurfaces for optical signal processing. *Phys. Rev. Lett.* **121**, 173004–173010 (2018).
41. Smolyaninov, A., El Amili, A., Vallini, F., Pappert, S. & Fainman, Y. Programmable plasmonic phase modulation of free-space wavefronts at gigahertz rates. *Nat. Photon.* **13**, 431–435 (2019).
42. Khoram, E. et al. Nanophotonic media for artificial neural inference. *Photon. Res.* **7**, 823–827 (2019).

43. Grigorescu, S., Trasnea, B., Cocias, T. & Macesanu, G. A Survey of Deep Learning Techniques for Autonomous Driving. *arXiv*:1910.07738 (2019).
44. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
45. NVIDIA TITAN RTX. https://www.nvidia.com/en-us/deep-learning-ai/products/titan-rtx/
46. Japan Captures TOP500 Crown with Arm-Powered Supercomputer. https://top500.org/news/japan-captures-top500-crown-arm-powered-supercomputer/ (2020)
47. Wang, T. Y. et al. An optical neural network using less than 1 photon per multiplication. *arXiv*:2104.13467 (2021).
48. Miscuglio, M. et al. Massively parallel amplitude-only Fourier neural network. *Optica* **7**, 1812–1819 (2020).
49. Zhou, T. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photon.* **15**, 367–373 (2021).
50. Zhou, C., Liu, L., Li, G. & Ying, Y. Modified engagement method for matrix operation. *Appl. Opt.* **34**, 7608–7614 (1995).

Figures

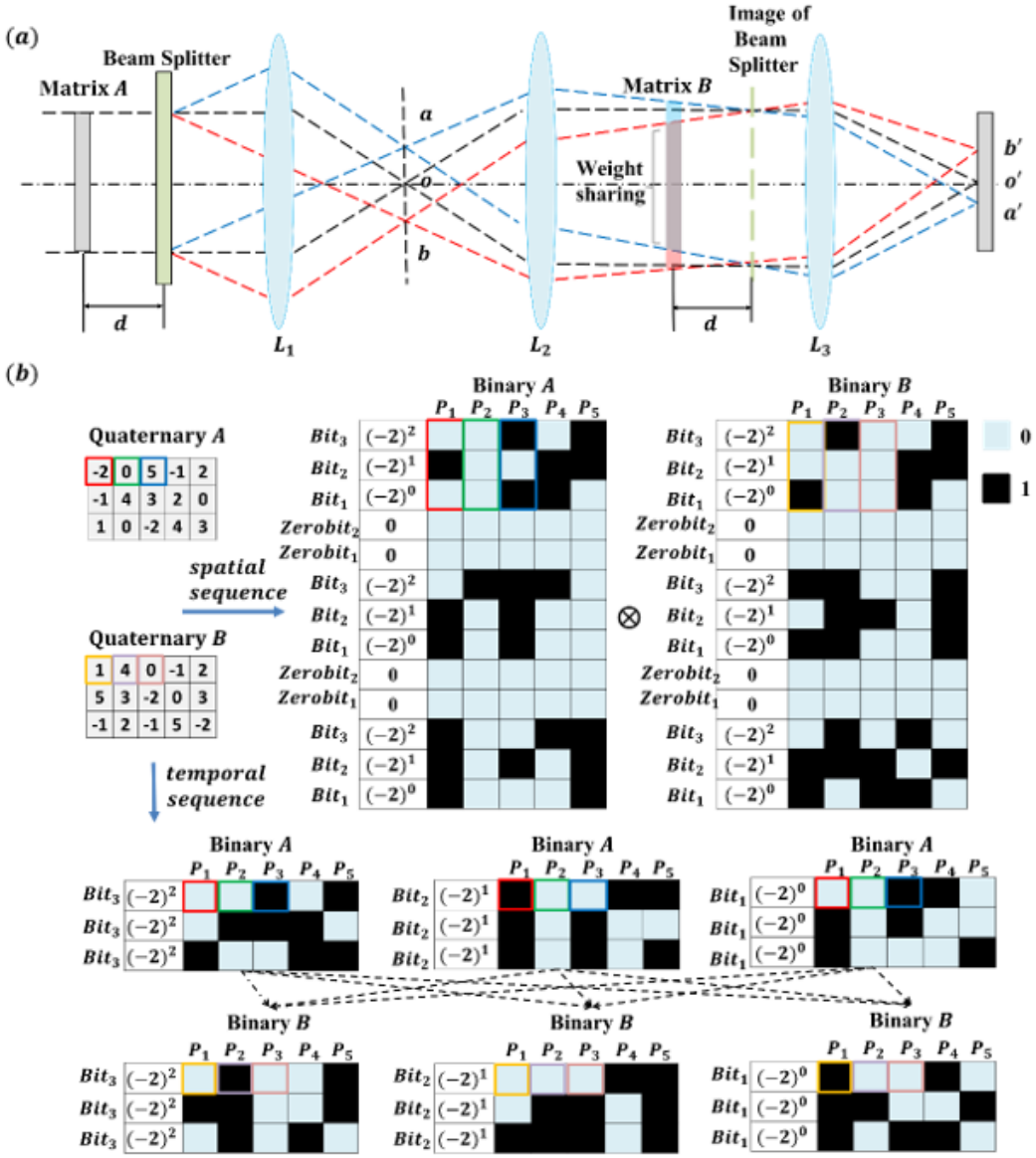

Figure 1

Schematic of the optical convolution accelerator based on OMica architecture: (a) Optical principle of OMica architecture. (b) Procedure of encoding the quaternary (2-bit) matrix into a binary (1-bit) matrix in two different modes—a spatial and temporal sequence.

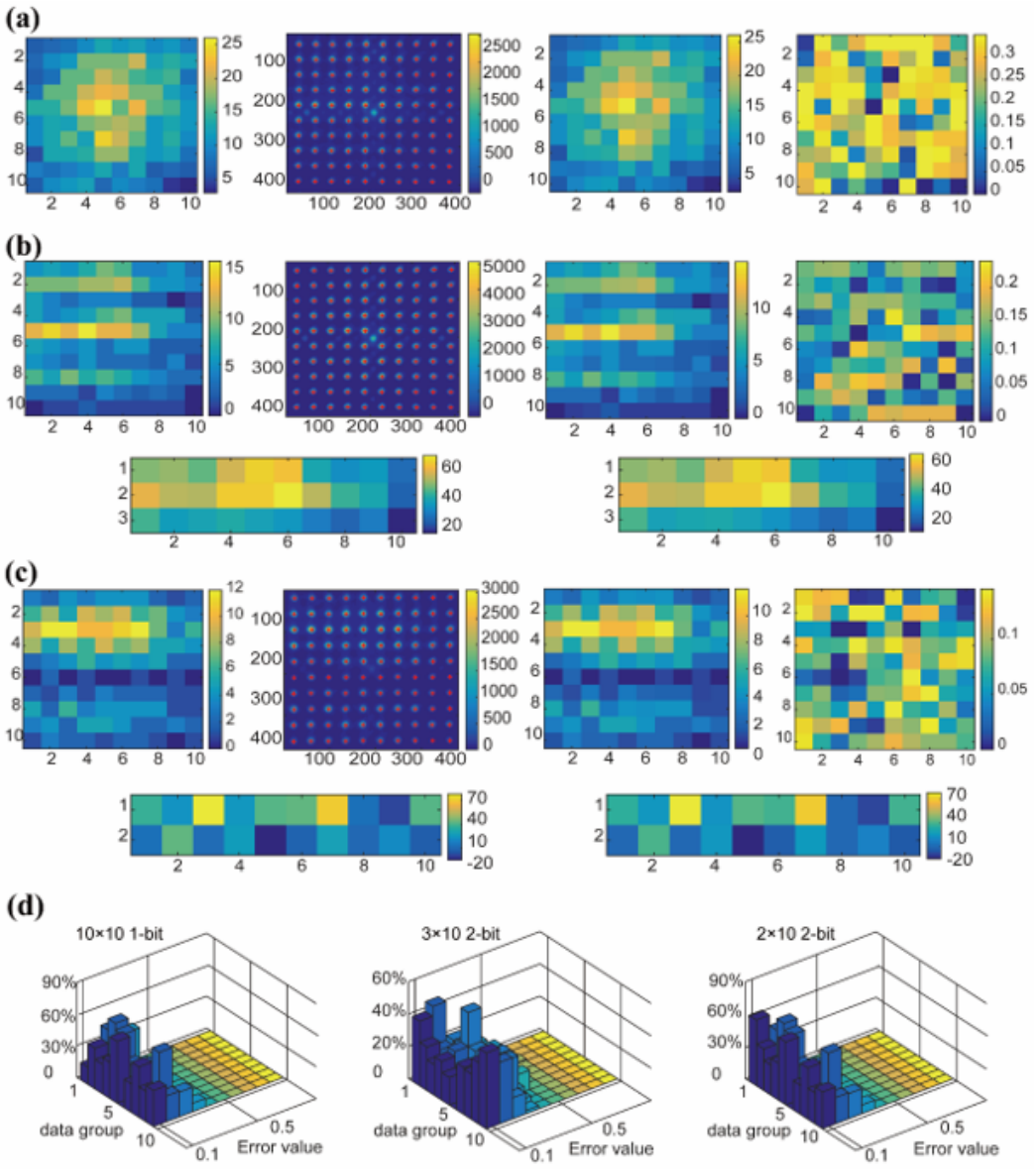

Figure 2

Experimental results of hybrid analog–digital matrix convolution for three groups of matrices based on spatial sequence encoding. (a) 10 × 10 binary (1-bit) matrices: the subfigures from left to right are the theoretical convolution values, light intensity distribution of the spot array denoting the convolution, experimental convolution results, and error map between the theoretical and experimental results, respectively. (b) 3 × 10 2-bit matrices: the subfigures from left to right are the theoretical convolution values before decoding, light intensity distribution of the spot array denoting the convolution, experimental convolution results before decoding, and error map between the theoretical and experimental results; the subfigures in the second row are the theoretical original convolution matrix and experimental convolution results after decoding, respectively. (c) 2 × 10 2-bit encoding matrices with negative elements: the subfigures in the first row from left to right are the theoretical convolution values before decoding, light intensity distribution of the spot array denoting the convolution, experimental convolution results before decoding, and error map comparing the theoretical and experimental results; the subfigures in the second row are the theoretical original convolution matrix and experimental convolution results

after decoding, respectively. (d) The error distribution of these three groups of data, from left to right, corresponding to 10 × 10 1-bit, 3 × 10 2-bit, and 2 × 10 2-bit encoding matrices with negative elements.

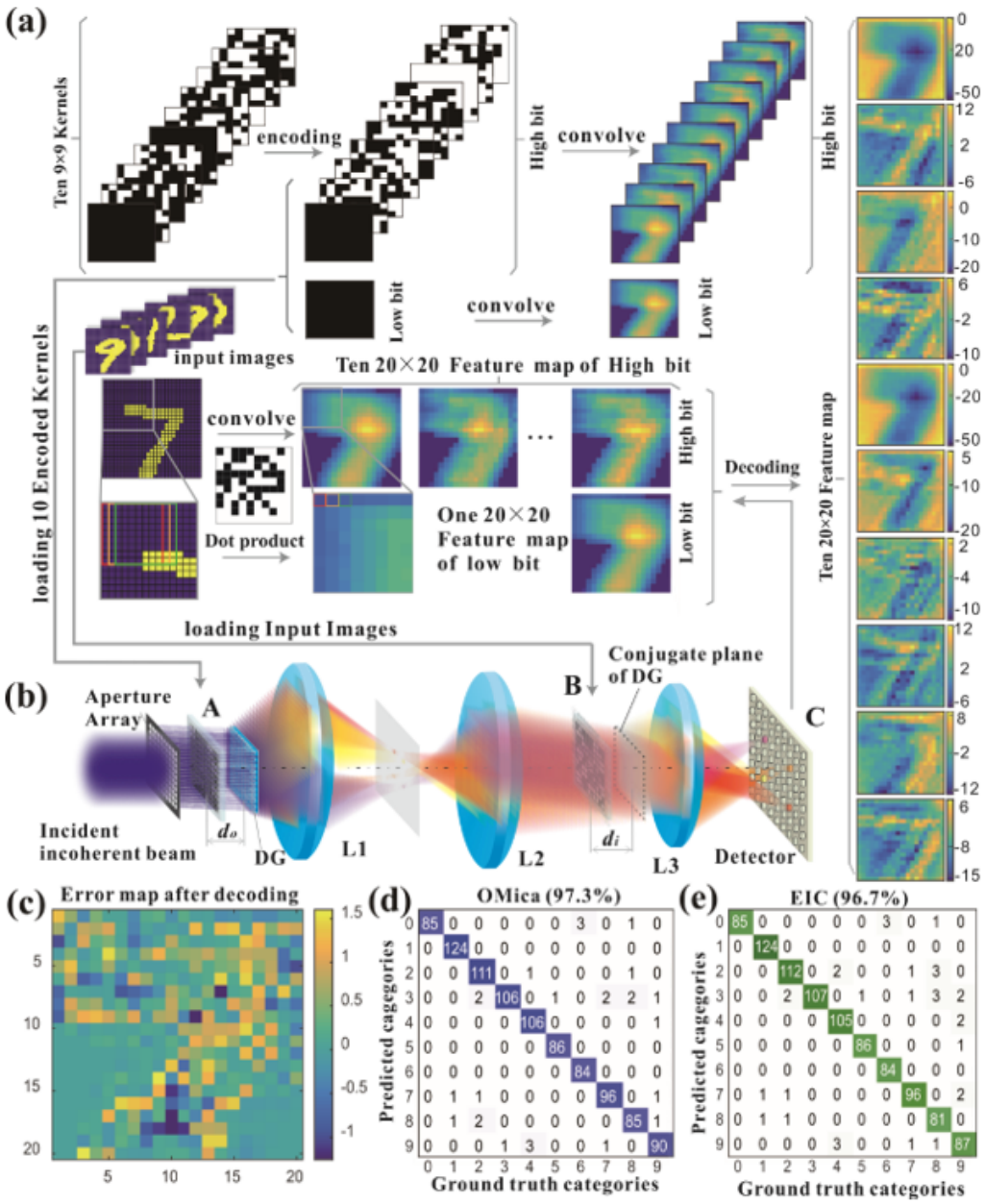

Figure 3

Optical convolution computing system realizes the convolutional neural network based on the MNIST dataset. (a) Execution of convolution operation by encoding each original convolution kernel into high-bit kernels and a low-bit kernel; (b) schematic of OMica architecture performing CNN inference; (c) the absolute error map comparing the theoretical and experimental results of the convolution of a handwritten digit 7 as an input; the confusion matrix of blind-testing 1000 images from the MNIST dataset when matrix convolutions are executed by optical hardware; (d) indicating a recognition accuracy of 97.3%; and (e) by pure electric hardware, indicating a recognition accuracy of 96.7%.

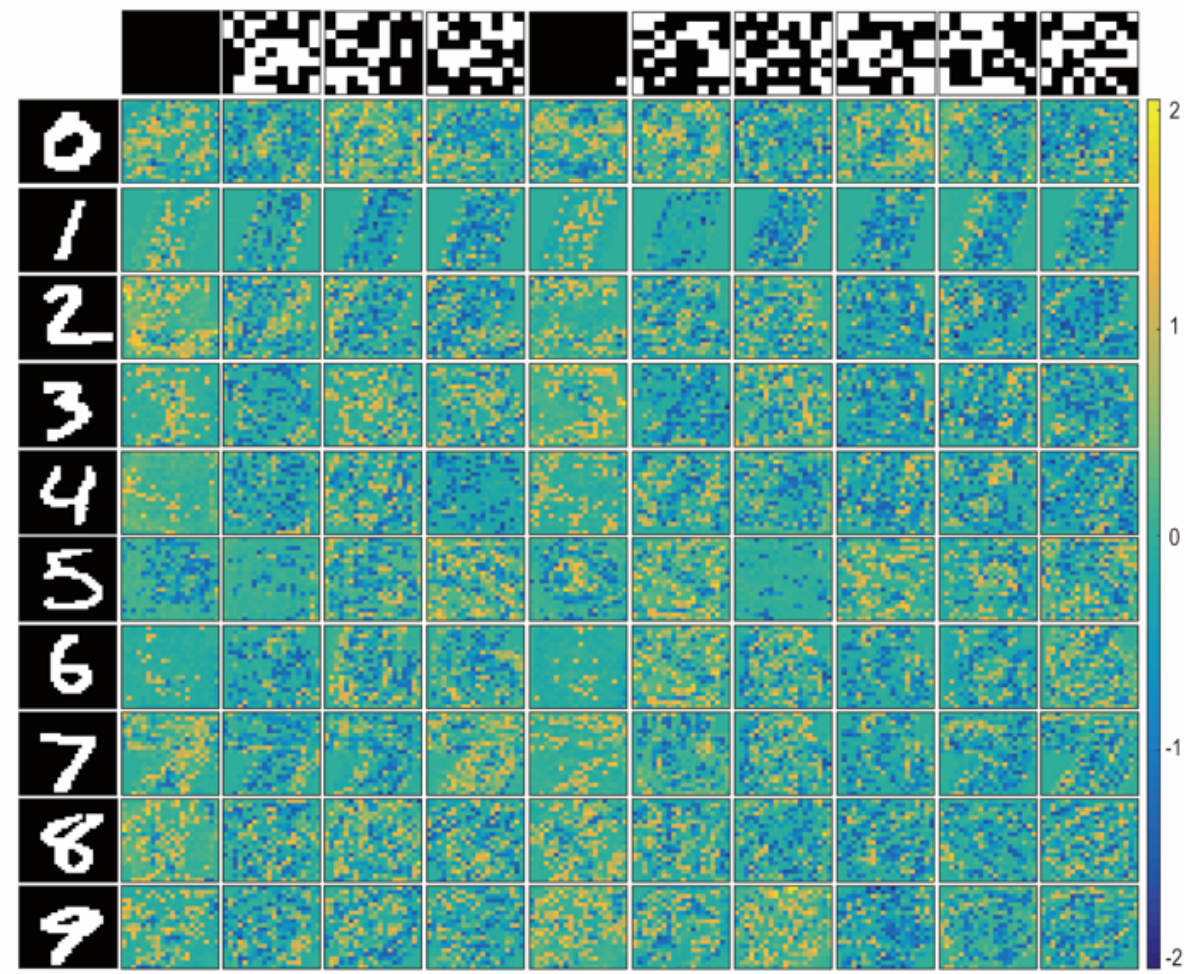

Figure 4

Typical error maps of convolution results between different input handwritten digits—from 0 to 9—and these ten convolution kernels after encoding.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationOpticalMultiImagingCastingAcceleratorforFullyParallelUniversalConvolutionComputing.docx