Research Article

Programming nonlinear propagation for efficient optical learning machines

Ilker Oguz,[a,*] Jih-Liang Hsieh,[a] Niyazi Ulas Dinc,[a] Uğur Teğin,[a,b] Mustafa Yildirim,[a] Carlo Gigli,[a] Christophe Moser,[a] and Demetri Psaltis[a]

[a]École Polytechnique Fédérale de Lausanne, Institute of Electrical and Micro Engineering, Ecublens, Switzerland
[b]Koc University, Department of Electrical and Electronics Engineering, Istanbul, Turkey

Abstract. The ever-increasing demand for training and inferring with larger machine-learning models requires more efficient hardware solutions due to limitations such as power dissipation and scalability. Optics is a promising contender for providing lower power computation, since light propagation through a nonabsorbing medium is a lossless operation. However, to carry out useful and efficient computations with light, generating and controlling nonlinearity optically is a necessity that is still elusive. Multimode fibers (MMFs) have been shown that they can provide nonlinear effects with microwatts of average power while maintaining parallelism and low loss. We propose an optical neural network architecture that performs nonlinear optical computation by controlling the propagation of ultrashort pulses in MMF by wavefront shaping. With a surrogate model, optimal sets of parameters are found to program this optical computer for different tasks with minimal utilization of an electronic computer. We show a remarkable decrease of 97% in the number of model parameters, which leads to an overall 99% digital operation reduction compared to an equivalently performing digital neural network. We further demonstrate that a fully optical implementation can also be performed with competitive accuracies.

Keywords: neural networks; nonlinear optics; fiber optics; surrogate optimization; neuromorphic computing; wavefront shaping.

1 Introduction

Machine-learning architectures have come to be dominated by artificial neural networks (ANNs). There are several reasons why this architecture is used so broadly. Initially, their similarity with biological neural networks[1] provided strong motivation to explore ANNs. At the same time, the fact that ANNs are universal machines[2] that are able to approximate any function breeds confidence that ANNs can carry out useful and difficult tasks. Perhaps most significantly, the fact that error backpropagation[3] has proven very effective in training such networks catapulted their application to a wide variety of problems. Ever-larger networks[4] have been adopted for tackling challenging tasks.[5] Empirically, it has been found that larger networks tend to perform better given a sufficiently large database of training examples. This has led to a "bigger is better" mentality.[6] However, the disadvantage of this mentality is the energy required to train and use very large networks. For instance, only training the language model GPT-3, which has 175 billion parameters, consumed 1.3 GWh of electricity, which is the energy required to fully charge 13,000 Tesla Model S cars.[7] Optics can help overcome this downside, since light propagation through a non-absorbing, nonscattering medium is a lossless linear operation.

Several approaches have been reported for the optical realization of ANNs. Wavefront shaping by diffractive surfaces or modulators, followed by propagation can implement ANNs and perform different tasks, such as classification and imaging.[8–12] Silicon photonics technology also allows the realization of reconfigurable optical computing with much smaller dimensions. Control structures such as Mach–Zehnder modulators[13] and microring resonators[14] can precisely manipulate light inside waveguides as building blocks of ANNs, performing operations such as multiplication or addition. However, their constraint to two dimensions diminishes the intrinsic three-dimensional scalability of optics.[15] Additionally, in both of those domains, nonlinearity is obtained generally by optoelectronic devices, which impairs energy efficiency and speed.[16]

Moreover, as optical ANNs are analog systems, their training is cumbersome. The gradient descent optimization algorithms can be applied *ex situ*[8] or *in situ*;[17] however, the system should be completely calibrated, or perfectly modeled. As an alternative, stochastic or deterministic gradient-free optimization algorithms could be utilized for the training. Stochastic, evolutionary algorithms, such as genetic algorithms, found themselves implementations[18,19], as they do not require any model of the systems and through efficient trial and error mechanisms, they can directly optimize the NNs with experiments, however, they require a very high number of trials.[20,21] On the other hand, deterministic, surrogate-based optimization algorithms are guided by a simple metamodel between input parameters and loss function directly and are proven to be able to provide excellent results in digital NNs with a smaller number of iterations.[22–24]

In contrast to determination and rigorous mapping of individual weights of ANN models to photonics devices, ANNs based on high-dimensional, fixed, and nonlinear connections, such as reservoir computers[25] or extreme learning machines[26] can directly be implemented with diverse optical dynamics, such as multimodal speckle formation,[27,28] random scattering,[29] and transient responses.[30] The necessary nonlinearity between neurons could be introduced by electronic feedback,[31] optoelectronic conversion,[29] saturable absorption,[32] and second-harmonic generation.[33] Moreover, the Kerr effect inside both single-mode[34] and multimode[35] fibers (MMFs) was shown recently to be an effective dynamic for realizing computing systems. Another set of studies showed also that the nonlinear interactions inside MMFs are tunable by wavefront shaping.[36–38] However, the configurability of this physical computing approach has been limited to the digital readout weights (RWs) to map the output of the transform to a desired inference,[25] even though it has been shown computationally that evolutionary algorithms could be used to select the initial fixed linear transform.[39]

In this paper, we present an optical ANN architecture that combines a relatively small number of digitally implemented parameters to control the very complex spatiotemporal transformation realized by optical wave propagation. Our experimental studies show that with a spatial light modulator (SLM), by simultaneously shaping light with data and a fixed programming pattern, we can induce the nonlinear transformation inside the MMF to perform desired computations. We find that the optimization of a small number of programming parameters (PPs), around 50 in our experiments, results in a remarkable performance of the optical computer. For instance, we will show that the system with ~2000 total parameters (TPs), PPs, and RWs combined, performs as well as a digital ANN with over 400,000 parameters for the face classification task on the CelebA[40] dataset (Table 2). Moreover, we demonstrate that the same method can be used to program the propagation inside MMFs to perform all-optical classification without the digital readout stage. In this case, the classification can be directly read out with a simple beam location sensor, further decreasing the number of TPs.

2 Methods and Results

2.1 Nonlinear Fiber Propagation and Its Programming for Higher Classification Performance

Our method consists of a nonlinear optical transformation and a programming algorithm. In this study, the optical experiment is selected to be the propagation of spatially modulated laser pulses through an MMF. The nonlinear propagation of an ultrashort pulse inside an MMF is a highly complex process that entails spatial and temporal interactions of electromagnetic waves coupled with hundreds of different propagation modes. In addition to light transfer between different channels due to linear mechanisms, such as bending and defects in the fiber, at high intensity levels light–matter interactions start to occur, and this results in energy exchange between different modes depending on products of these modes. These interactions can be concisely described with the multimode generalized nonlinear Schrödinger equation,[41]

$$\frac{\partial A_p}{\partial z} = \underbrace{i\delta\beta_0^p A_p - i\delta\beta_1^p \frac{\partial A_p}{\partial t} - i\frac{\beta_2^p}{2}\frac{\partial^2 A_p}{\partial t^2}}_{\text{dispersion}} + \underbrace{i\sum_n C_{p,n}A_n}_{\text{linear mode coupling}}$$
$$+ \underbrace{i\frac{n_2\omega_0}{S_p}\sum_{l,m,n}\eta_{p,l,m,n}A_lA_mA_n^*}_{\text{nonlinear mode coupling}}. \tag{1}$$

Here A_p is the complex coefficient of the p'th normalized mode, where β_k^p is the k'th order propagation constant for this mode. $C_{p,n}$ is the linear coupling coefficient between modes p and n, which becomes nonzero when there are nonidealities in the waveguiding structures, such as ellipticity, bending, or impurities. n_2 is the nonlinear refractive index of the material, ω_0 is the center angular frequency, and S_p is the effective mode area of mode p in the fiber. $\eta_{p,l,m,n}$ is the nonlinear coupling coefficient between modes; it mainly depends on the similarity between spatial shapes of modes (F_p) and can be calculated as $\eta_{p,l,m,n} = \int dx dy F_p^* F_l F_m F_n^*$. Due to the short propagation length in the study, the higher-order dispersion effects are not shown. Similarly, Raman scattering is not observed dominantly in the experiments (see Fig. 6); hence in the nonlinear mode coupling part, only the contribution of the Kerr effect is shown.

The triple multiplications ($A_lA_mA_n^*$) with varying strengths of interaction ($\eta_{p,l,m,n}$) between different sets of modes demonstrate the nonlinear and immensely multimodal aspect of the interactions such that modeling the transform of a single pulse on the setup provided in Fig. 1 would take 50 min with a graphics processing unit (GPU).[42] The physical optical system carries out this complex spatiotemporal transformation "effortlessly." The transformation is programmed with a relatively small number of PPs. PPs are selected to customize the MMF processor for the specific task by wavefront shaping, controlling the optical power, and the placement of the data and diffraction angle on the SLM. This way the PPs can modify implicitly the distribution of A_p, which determines the nonlinear computation the MMF performs on the data. For instance, changing the diffraction angle on the SLM away from the optical axis moves the beam's focus from the center of the MMF, which is placed in the focal plane of the coupling lens. This primarily excites higher-order modes and effectively changes the part of $\eta_{p,l,m,n}$ tensor that acts on the data. The combination of wavefront shaping and data encoding as shown in Fig. 1 can be formalized as follows, where the complex encoded value on the SLM $E_{\text{SLM}}^k(x,y)$ is given by

$$E_{\text{SLM}}^k(x,y) = \text{WF}(x,y)\exp(iD^k(x,y))$$
$$= |\text{WF}(x,y)|\exp(i(D^k(x,y) + \text{Arg}(\text{WF}(x,y)))). \tag{2}$$
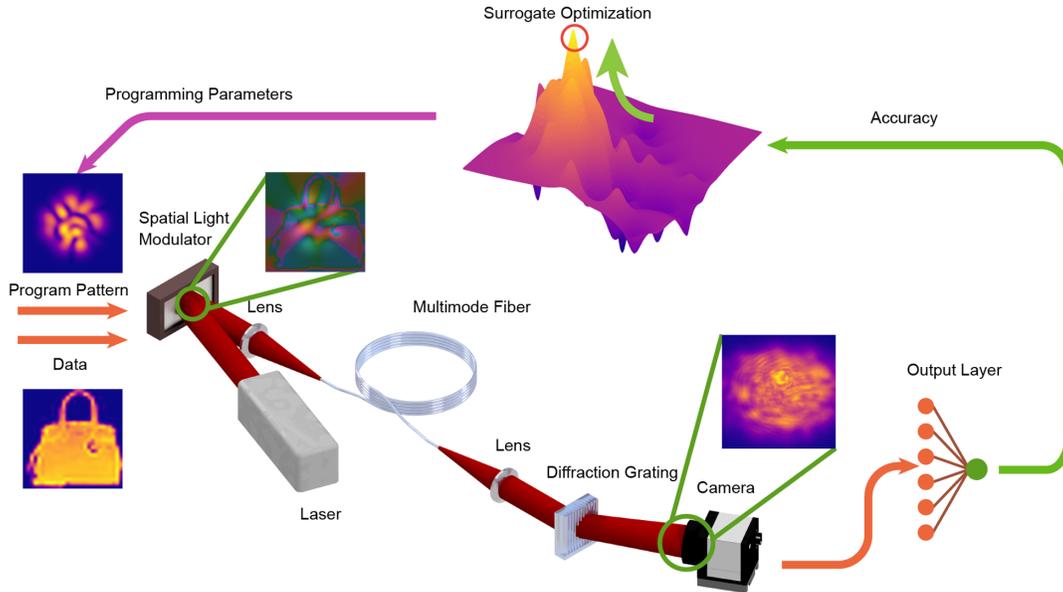
Fig. 1 Experiment flow for programming optical propagation for a computational task. The SLM modulates the laser pulses with the data sample overlaid with a fixed programming pattern calculated by the programming patterns. The beam is coupled to an MMF; the pattern after propagation is recorded with a camera. A trainable output classification layer calculates the task accuracy, which is fed back to the surrogate optimization algorithm. The algorithm improves the task performance by exploring different PPs and refining potential solutions.

The amplitude modulation with the phase-only SLM is realized by modifying the strength of a blazed grating (see Method 3 in the Supplementary Material), whereas the data $[D^k(x, y)]$ are encoded as a phase pattern. The wavefront controlling shape $[\mathrm{WF}(x, y)]$ is a complex combination of N different linearly polarized fiber modes $F_n(x, y)$, and the coefficient of each mode is controlled by two parameters, for real and imaginary parts. Even though other bases for controlling the wavefront can also be effective, we used the propagation modes of the optical fiber, $F_n(x, y)$, since they are interpretable, orthonormal, and guaranteed to be within the fiber's numerical aperture. The portion of the wavefront shape that contains the $2N$ PPs (a_n) is

$$\mathrm{WF}(x, y) = \sum_{n=1}^{N} (a_{2n-1} + i a_{2n}) F_n(x, y). \qquad (3)$$

During each step of the programming procedure, the optical system processes the dataset for a set of values for PPs, and the task performance is measured. Depending on the task, the performance metric could be the training accuracy of the final RWs (Fig. 2), or the ratio of correctly placed outputs for all-optical tasks (Fig. 3). Throughout the process of programming optical system, a surrogate optimization algorithm[22] selects values of PPs to explore the dependency between them and the loss function, which is the negative of the accuracy in classification tasks, and finally finds the globally optimal set of PP values for the best computational performance. The same procedure is also widely used in the optimization of NN architectures;[43] its details are provided in the Appendix.

In the experimental realization of the method, a commercially available mode-locked laser at 1030 nm (Amplitude Laser, Satsuma) with 125 kHz repetition rate is used, and the pulse length is set to be the longest possible, 10 ps, with an internal dispersive grating stretcher for obtaining a longer dispersion length. This maximizes the length in which nonlinear interactions effectively occur in the 5 m-long graded-index multimode fiber (OFS, bend-insensitive OM2, 50 μm core diameter, 0.20 NA, 240 modes at 1030 nm), without being hindered by dispersion-induced pulse broadening. Before coupling to the MMF, the laser beam is sent onto a reflective phase-only SLM (Meadowlark HSP1920), which encodes the data to be transformed by the optical system to 0 to 2π phase retardation at each pixel location and combines them with a complex pattern containing the PPs, as described in Eq. (1). The intensity of the coupled pulsed laser beam is also treated as one of the PPs; it is controlled via a half-wave plate mounted on a motorized rotation stage followed by a polarizing beam splitter and is optimized through the surrogate model. Once the PPs are determined, the data portion of the modulation changes with every sample, while the programming part stays the same. Upon exiting the fiber, the beam is collimated with a lens and sent to a blazed grating. The dispersion due to the grating leads to a camera recording, in which both spatial and temporal characteristics of the output beam are present.

The classification accuracies are calculated by training a simple regularized linear regression algorithm with L2 regularization on the pixel intensity values of the recorded images on the camera in Fig. 1. The linear regression maps the recorded image to classification results by pointwise multiplication with RWs and summation. Finally, the pairs of PPs and corresponding task performance of the system are supplied to the surrogate algorithm that optimizes the optical transform. After acquiring the performance metric for different sets of PPs, the surrogate

optimization algorithm creates mappings between the performance of the system and any given set of PPs. The surrogate algorithm continuously refines the model and increases the performance on the task at the same time. Video 1 and Fig. 2 illustrate such an experiment, where the data transform with the MMF was programmed for a higher classification accuracy on a small subset (2%) of the Fashion-MNIST dataset, which consists of 1200 training images and 300 test images of 10 different classes of fashion items.[44]

As shown in Fig. 2, for the first 105 iterations, the surrogate optimization algorithm broadly samples the parameter space to create the initial mapping between PPs and the performance metric. After this phase, an area with the potential of yielding the best result is selected and sampled in finer steps. Gradually,

the changes become smaller, and the algorithm converges to a solution. Figure 2(b) provides a closer look at the progression of the process, where each data point represents an iteration, the two-dimensional random projection of 46 wavefront shaping parameters versus the training accuracy. The initial homogeneous sampling of the parameter space and final fine-tuning can be observed. Similarly, Fig. 2(c) shows that after exploring various levels of optical intensity, hence nonlinearity, the convergence led to a higher light intensity for obtaining a more efficient nonlinear optical transform. Even after converging, the search algorithm probed different intensity levels to ensure not being stuck in a local minimum but came back to the same level confirming it is the optimum. This is also made possible by converging to a preferred oblique excitation of the fiber,
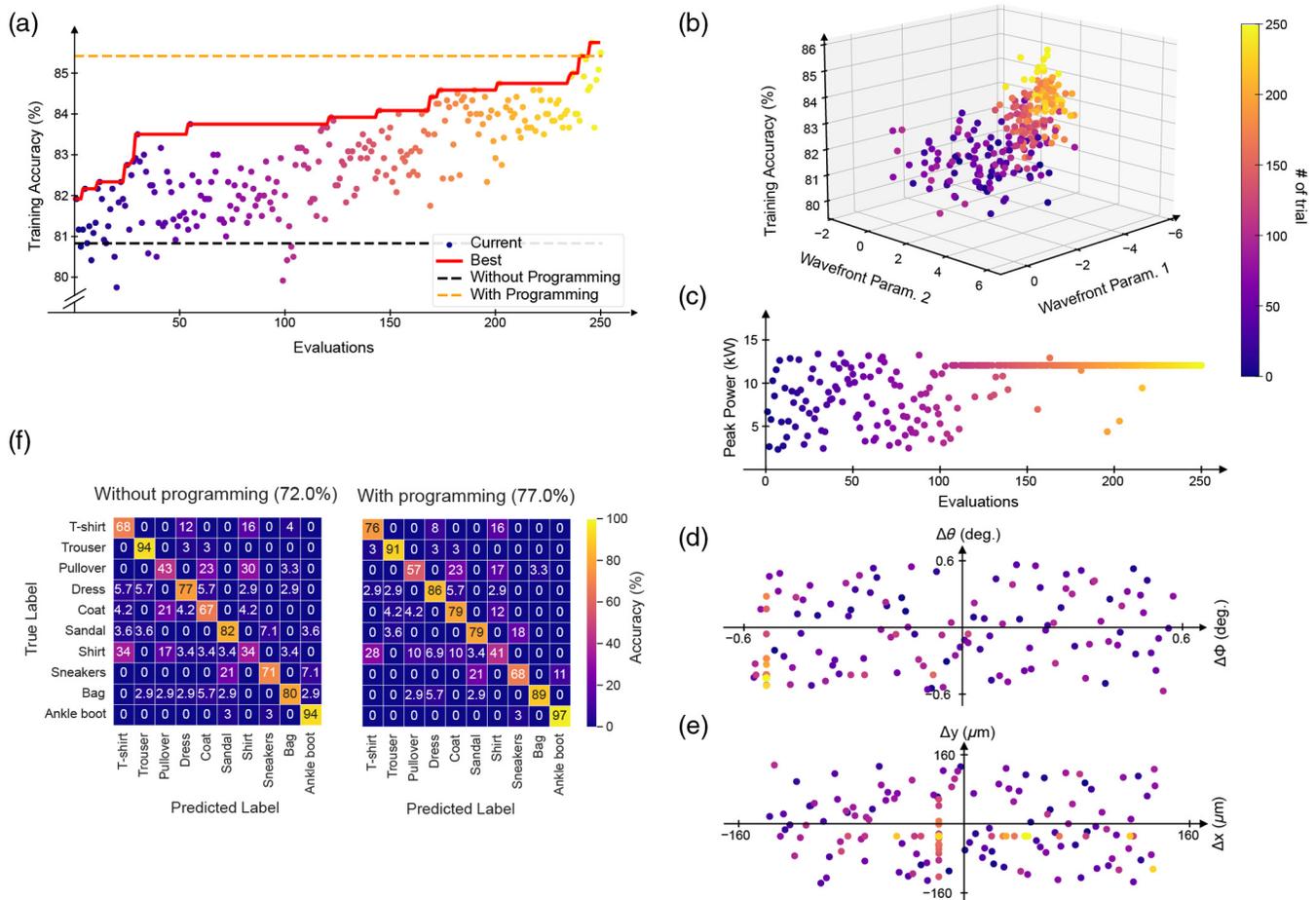
Fig. 2 Programming the MMF propagation for higher classification performance on Fashion-MNIST dataset. (a) Training accuracy during the progress of the programming procedure. The horizontal line labeled "without programming" shows the accuracy level when PPs are set to zero and "with programming" indicates the level when the PPs found by the programming algorithm are used. The colors of circles indicate their sequence in the training. (b) Relation between wavefront shaping parameters and training accuracy. Forty-six different wavefront shaping parameters are shown in two dimensions by means of random projection into two dimensions for visibility. (c) Peak power of pulses during the programming procedure. (d) Change of the diffraction angle on the SLM in horizontal and vertical directions ($\Delta\phi$ and $\Delta\theta$). (e) Shift of image on the SLM in horizontal and vertical directions (Δx and Δy). (f) Confusion matrix and average accuracy on the test set, without and with the programming of the transform (Video 1, MP4, 2.31 MB [URL: https://doi.org/10.1117/1.AP.6.1.016002.s1]).

as shown in Fig. 2(d), allowing for stronger coupling to higher-order fiber modes as well, hence benefiting from the multimodality of the fiber. Overall, programming the optical propagation by optimized PPs improved the classification accuracy both on the training and the test sets by about 5% compared to using a not programmed propagation[35] and reached 77% accuracy on the test set. In comparison, seven-layer digital convolutional neural network (CNN) LeNet-5[45] yields 77.9% accuracy when trained with the same dataset on a GPU (for details see Method 1 in the Supplementary Material). Later, we present another approach for programming the propagation, in which PPs are combined with data through convolution, and 79.0% test accuracy is reached on the same task.

2.2 All-Optical Computing with Propagation in Optical Fiber

In Fig. 2, the PPs were optimized to modify the optical transform inside the MMF to improve the performance of the combination of the optical system and the digital readout layer.

To further demonstrate the programming capacity of our approach, the inference on input samples was done all-optically without any RWs by only using the center location of the output speckles, as shown in Fig. 3. For a binary classification problem, the input is classified as either "0" or "1," depending on which side of the classification line the center of the output beam resides. Hence, only three parameters are required at the output for defining any line in two dimensions. Figure 3 illustrates the programming procedure for all-optical classification of the dataset, consisting of 1200 training and 300 test chest radiography images, equally sampled from patients with and without COVID-19 diagnosis.[46]

Before programming the system, a linear classifier received the distribution of center locations without any control patterns and drew a classification boundary between positive and negative samples. As the transform is random, the classifier could only produce training and test accuracies around 50%. Then the decision boundary is kept the same, and the training accuracy is improved by optimizing the PPs on the SLM with the surrogate model to separate the center location distributions for samples
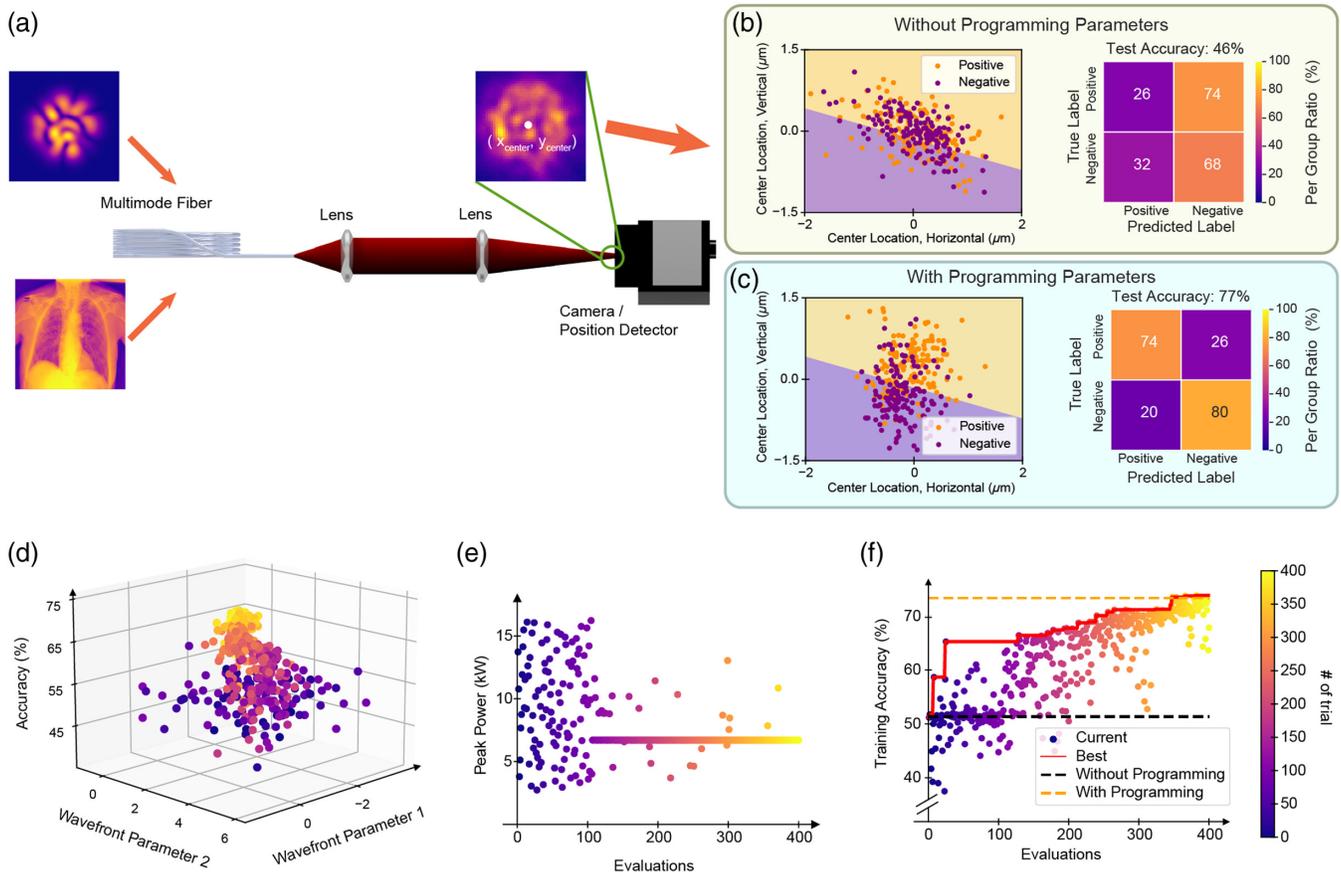
Fig. 3 Programming procedure for all-optical classification of chest radiographs. (a) The schematic of the experiment, the data, and the control pattern are sent together to the SLM, and the fiber output pattern is imaged onto a camera. (b), (c) Distribution of the beam center locations and corresponding confusion matrices for the test set, without and with the programming of the transform. (d) Distribution of training accuracies with respect to the selection of wavefront shaping parameters. (e) Selected power levels for each iteration of the programming procedure. (f) Progression of training accuracy during training. The color map relates the color of circles to their sequence in the training, and it applies to (d)–(f) (Video 2, MP4, 4.17 MB [URL: https://doi.org/10.1117/1.AP.6.1.016002.s2]).

Table 1 Comparison between neural networks and all-optical classification system.

Network structure	Total number of parameters	Operations per sample on digital computer (FLOP)	Accuracy on melanoma dataset (%)	Accuracy on COVID-19 dataset (%)
LeNet-5	61,026	844,520	63.9 ± 1.6	73.2 ± 2.4
EfficientNetB6	4.1×10^7	4.2×10^8	77.3 ± 0.6	88.3 ± 0.6
MMF + classification with output location (with programming)	55	2029	61.3	77.0

with positive and negative labels, as shown in Video 2. This procedure improved the accuracy on the test set from 46% to 77%, as shown in Table 1. This performance, realized all-optically with only 55 TPs, compares favorably to LeNet-5, which uses about 61,000 parameters. A state-of-the-art, pretrained ANN (EfficientNetB6[47]) with 4.1×10^7 parameters achieves an 88.3% test accuracy when fine-tuned for the same task. Similarly, on the task of classifying skin lesions of equally sampled benign (nevus) and malignant (melanoma) case images,[48,49] 1200 samples being training and 300 test, all-optical system yields a 61.3% test accuracy.

2.3 Different Wavefront Shaping Approaches for Programming the Optical Transform

For the two experiments shown previously, the optical transform was programmed through the multiplication of fields by encoding data and PPs [i.e., D and WF terms in Eq. (2), respectively] as shown in Fig. 2 and detailed in Fig. 4(a). In this section, we demonstrate that the transform can be achieved with two additional ways of wavefront shaping, as depicted in Fig. 4.

For the modification of the phase, the control pattern was again formed, as described in Eq. (2). Thus this function is elementwise digitally multiplied with the input image from the dataset and placed on the SLM as visualized in Fig. 4(f). For the N'th sample of the dataset, the field diffracted by the SLM becomes $E^N_{SLM}(x, y) = \exp(i(D^N(x,y)\text{Arg}(\text{WF}(x,y))))$. After optimization of the PPs, the test accuracy on the subset of Fashion-MNIST reached 78%.

Alternatively, convolutional filters can be used to amplify or attenuate different parts of the angular spectrum of the field—hence its mode decomposition inside the MMF. Importantly, convolutional filters can be applied fully optically by filtering in the Fourier plane.[50] Figure 4(k) depicts that convolution can also program the nonlinear propagation and could reach 79% test accuracy on the same dataset when each element of the convolution kernel is set as PPs. Hence, the $c \times c$ convolution kernel could be written as $A = \begin{bmatrix} a_1 & \cdots & a_c \\ \vdots & \ddots & \vdots \\ a_{c^2-c+1} & \cdots & a_{c^2} \end{bmatrix}$ in terms of programming c^2 PPs. Then the field modulated by the convolution filtered N'th sample of the dataset is $E^N_{SLM}(x, y) = \exp(i(A \circledast D^N(x,y)))$. Similarly, on the 1200 training and 300 test samples from the MNIST-digits dataset 94.3% accuracy, which is comparable to the 94.9% accuracy of a nine-layer digital ANN with ~420,000 parameters, could be reached with the same approach, demonstrating that different wavefront shaping strategies could realize the enhanced interactions within the optical fiber.

2.4 Transferring Programming Parameters across Different Tasks and Datasets

The optimization of PPs from scratch requires processing the selected dataset on the experimental system more than 100 times while modifying the PPs. With ~50 frames per second rate, this process takes a few hours for a dataset of 1500 images. In addition to switching to faster optoelectrical devices, the ability to transfer previously optimized PPs to new tasks or datasets would boost the practical utility of this approach as a general-purpose component that can be quickly deployed on different problems. This ability is first demonstrated with the reusability of PPs on different tasks for the same dataset. After finding the set of optimal PPs for the task of classifying the gender of the person in an image from Celebrity Face Attributes dataset (CelebA),[40] the same set of PPs is used for determining the age of the person. The only training required for the transfer between tasks is the determination of RWs without any new surrogate optimizations involving the optical system. Table 2 compares the performance of the PP transfer between age and gender tasks by fully programming the system for each task separately, showing that the test accuracy with the parameter transfer follows the accuracy of programming from scratch. Without programming the fiber, the test accuracy on the age classification task is 59.0% and 2026 RWs are used. After optimizing additional 52 PPs (wavefront shaping and experimental parameters, reaching 2078 TPs), the test accuracy reaches 67.0%, performing better than a digital nine-layer CNN with about 412,000 parameters. When these optimized 52 PPs are used on the gender classification task on the same dataset only with retraining of the RWs, the accuracy on the new task is 76.0%, which is similar to the 76.3% achieved by programming the system from scratch with the gender database. The same findings hold true when the initial programming is done on the gender task and parameters are transferred to the age task.

Furthermore, we find that with transfer learning,[51] the optimized PPs can also be utilized with a different dataset after only a short corrective programming. The PPs optimized for the COVID-19 classification dataset with the all-optical approach are transferred to the task of classifying skin lesions between benign (nevus) and malignant (melanoma) case images[48] (Fig. 5). However, directly transferring the PPs from former to latter resulted in a test accuracy of 47.67%, which is similar to a random prediction. In corrective programming, a smaller set of parameters (11 in total) is designated for optimizing the previously acquired set of PPs. These 11 parameters are combined with 52 PPs by repetition of each element multiple times and element-wise addition. Thus an optimal set of PPs is found in the proximity of the initial expectation by optimizing in a lower dimensional search space. Decreasing the dimensionality
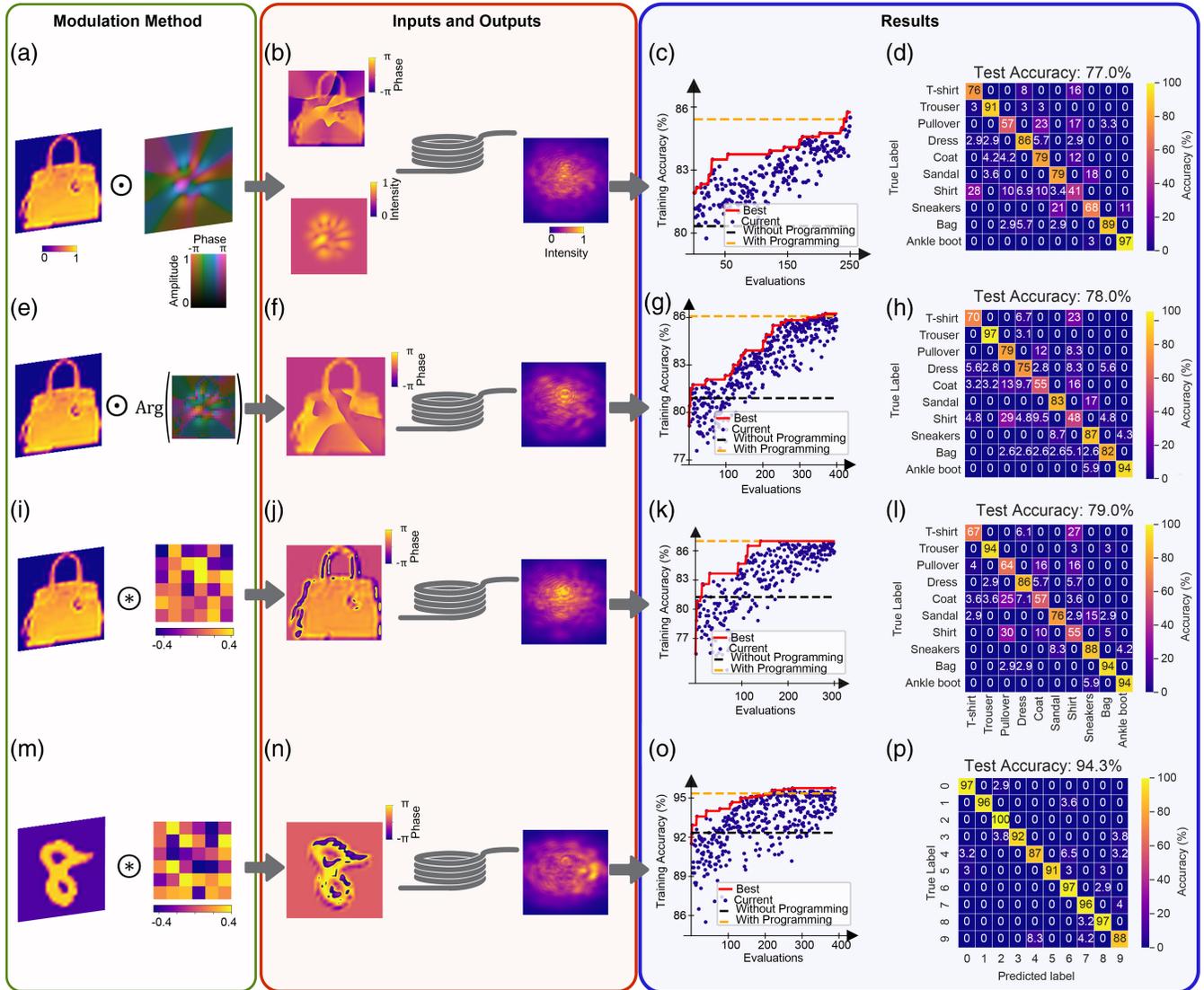
Fig. 4 Programming the optical transform using (a) phase addition and amplitude modulation, (e) multiplication with phase, and (i), (m) convolution. (b), (f), (j), (n) Example of programmed patterns on the SLM and recorded intensity patterns after the propagation inside the optical fiber for the given input pattern; for (f), (j), and (n), the intensity is not modulated. (c), (g), (k), (o) depict the progression of training accuracies during programming iterations. The confusion matrices on (d), (h), (l), (p) illustrate the classification performance of the programmed optical transform with different methods.

enables a convergence in fewer iterations. Compared to the complete programming of the system in 300 iterations, corrective programming starts from a similar initial accuracy, and after 80 iterations instead of 300, reaches the same final test accuracy.

2.5 Role of Nonlinearity in the Optical Transform

Nonlinear activation functions between linear mappings in NNs allow them to approximate complex, nonlinear functions. Without nonlinearities, an NN would be limited to representing only linear transformations of the input data, which would severely limit its ability to model real-world data and solve complex problems. Similarly, in the proposed method, nonlinearities play a crucial role for the generalization performance. In addition to fixed nonlinearities acting on the input data, such as

phase modulation and intensity detection, optical nonlinear effects provide controllable means to introduce nonlinearity to the information transform. One of the main factors affecting the extent of optical nonlinearity is the intensity level of the beam. In Fig. 6, the effect of the peak power of laser pulses on the optical and data processing characteristics of the experiment is analyzed. Since the pulse length and repetition rate of the laser are measured beforehand, the peak power level is calculated from the average laser power by dividing it by the pulse length and repetition rate. In accordance with Kerr nonlinear effects in graded-index MMFs, the spectra become broader with higher peak powers, and the intensity of the beam focalizes to the center due to Kerr beam self-cleaning. This also affects the performance of the computational task, and, as Fig. 6(a) depicts, up to a peak power level of 7 kW, increased nonlinearity

Table 2 Performance of different CNNs and optical computing methods on the CelebA dataset.

Network structure	Total number of parameters	Operations per sample on digital computer (FLOP)	Test accuracy on age task (%)	Test accuracy on gender task (%)
LeNet-1	1702	275,724	60.3 ± 1.0	70.7 ± 1.0
LeNet-5	61,026	844,520	63.8 ± 1.4	75.4 ± 2.1
Nine-layer convolutional NN	411,794	65,163,532	65.3 ± 0.1	80.1 ± 0.7
MMF + linear output layer	2026	4050	59.0	69.0
Programmed MMF for age task + linear output layer (trained for the corresponding test task)	2078	6075	67.0	76.0
Programmed MMF for gender task + linear output layer (trained for the corresponding test task)	2078	6075	64.7	76.3
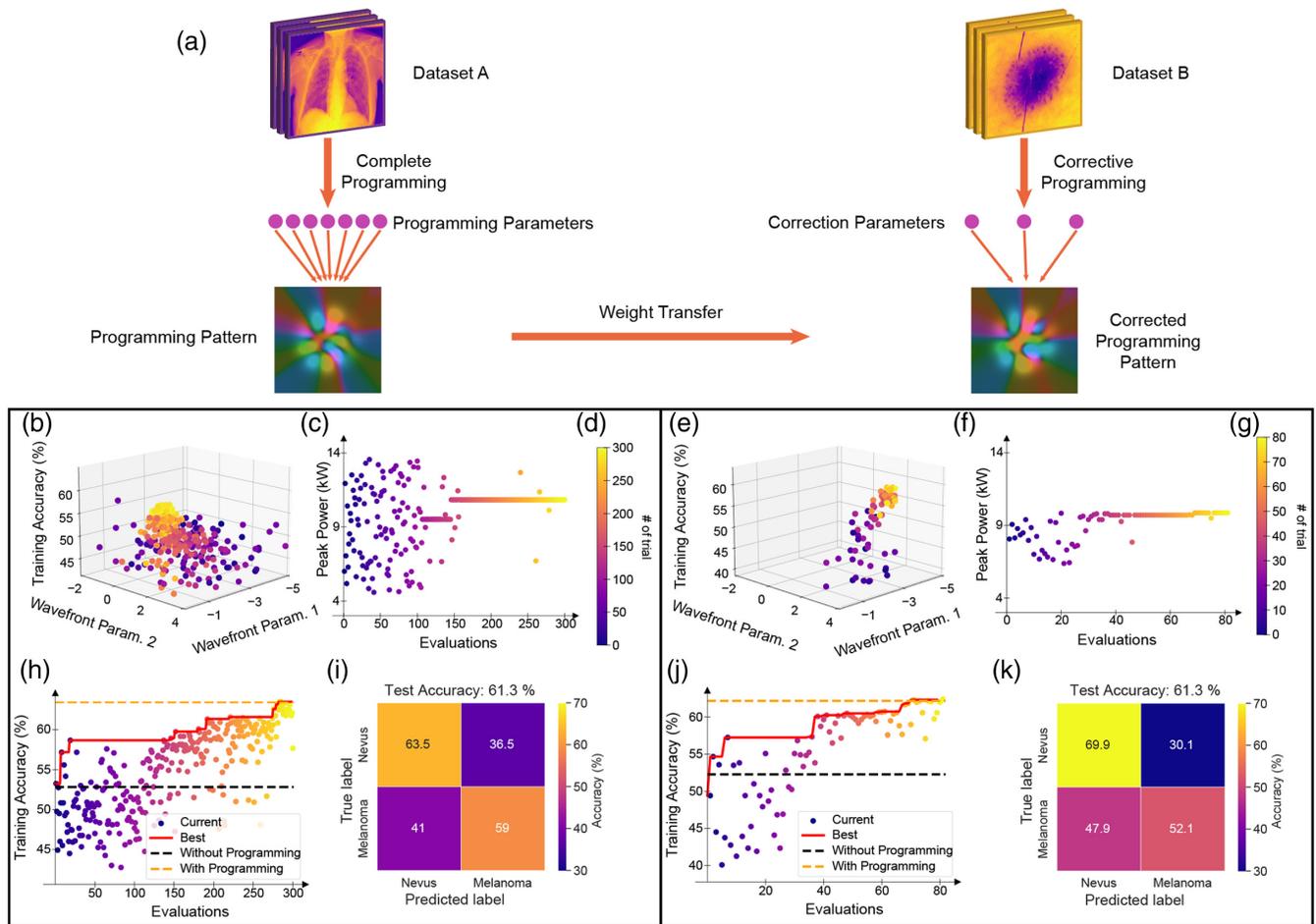
Fig. 5 Using previously dedicated parameters on a new dataset with corrective programming. (a) Procedure for transferring the PPs. (b)–(d), (h), (i) The experiment when the PPs are fully programmed without any prior knowledge. (e)–(g), (j), (k) Corrective programming of parameters. (b), (e) Relation between wavefront shaping parameters projected to two dimensions and the training accuracy. (c), (f) Peak power of pulses at the fiber entrance. (d), (g) Color bar for coding the iteration number related to each data point on (b), (c), (h) and (e), (f), (j). (h), (j) Training accuracy during the progress of the programming procedure. (i), (k) Confusion matrix and average accuracy on the test set.
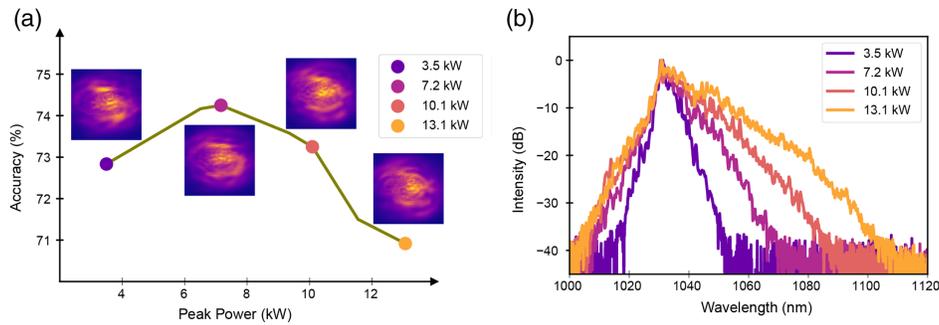
Fig. 6 Dependency of the training accuracy on the CelebA gender classification task, diffracted beam shape, and spectrum on the optical intensity level, with all other PPs set to zero. (a) Camera images for the same input image and the task accuracy for different pulse peak powers. (b) Optical spectrum after propagating in the fiber at different power levels for the same sample from the dataset.

improves accuracy. Above this value, performance monotonously decreases, possibly due to the deleterious nature of beam self-cleaning on the modal distribution. This process couples the energy from higher-order modes to the fundamental mode.[52] Depending on the original distribution of energy between propagation modes, the optimal power for the task can change. For instance, in contrast to the one in Fig. 6, the experiment in Fig. 2 utilized other PPs than the intensity level, and especially with oblique coupling, higher-order modes are excited more; hence, the optimal peak power level is found to be much higher, around 13 kW.

3 Discussion

3.1 Computation Speed and Energy

The speed of inferences is limited by the refresh rate of the liquid crystal SLM. This limitation can be overcome by switching to a faster wavefront shaping method, for instance, by utilizing commercial digital micromirror devices, which can reach 30,000 frames per second.[53] Since the number of modes in the MMF is much smaller than the number of pixels on commercial SLMs, different lines of the SLM could be scanned with the beam by a resonant mirror, allowing up to 25 million samples per second data input rate. Moreover, the fixed complex modulation or convolution operations can be implemented with optical phase masks, bringing the digital operation count further down. Similarly, instead of the digital readout layer, a broadband diffractive element can realize the linear projection step. As it is analyzed in Note 2 in the Supplementary Material and visualized in Fig. 7, implementing the same optical computer with a selection of commercially available, high-speed equipment such as digital micromirror devices and quadrant photodiodes, 25 TFLOP/s performance could be reached with a total power consumption of 12.6 W, which is significantly lower than 300 W consumption of a GPU with a comparable performance.[54]

3.2 Stability and Reproducibility

The reproducibility of experiments is crucial for consistent comparison between different sets of PPs during programming and long-term usability of determined PPs. To investigate
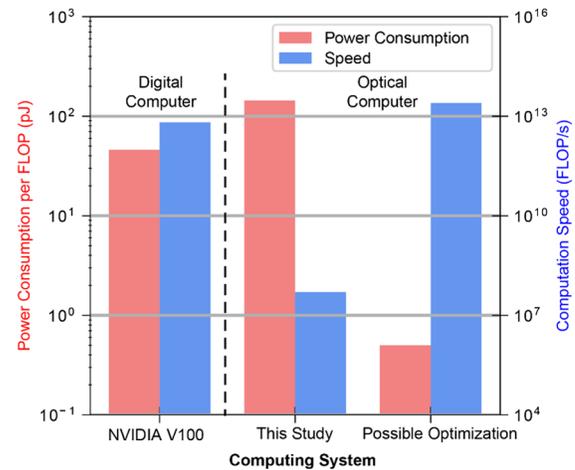
Fig. 7 Power efficiency and speed comparison between different computational approaches. The possible optimization refers to incorporating a digital micromirror device, a resonant mirror, and an optical phase mask in the optical computer.

reproducibility, the inference experiment is repeated every 5 min for the same PPs and RWs on the same task over 15 h. As shown in Fig. S1 of the Supplementary Material, the first and final test accuracies are the same, and the standard deviation of the test accuracy over time is 0.3%, indicating a very stable experimental inference.

In conclusion, programming nonlinear propagation inside MMFs with wavefront shaping techniques can exploit complex optical interactions for computation purposes and achieve results on par with multilayer neural networks while decreasing the number of parameters by more than 97% and potentially consuming orders of magnitude less energy for performing the equivalent number of computations. This shows the capacity of nonlinear optics for providing a solution for the exponentially increasing energy cost of the machine-learning algorithms. Not being limited to nonlinear optics, the presented framework could be used for efficiently programming different high-dimensional, nonlinear phenomena for performing machine-learning tasks.
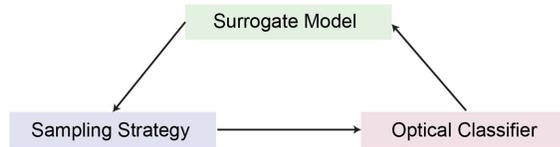
Fig. 8 Flowchart of the programming procedure.

4 Appendix: Programming Procedure

The optical experiment is considered as a whole with the final classifier, called the optical classifier in Fig. 8. For each set of PPs given by the sampling strategy, the optical classifier returns a training score to the surrogate model. First, the data are transformed by the optical system as detailed in Method 2 in the Supplementary Material. For the experiments that perform classification with grating-dispersed fiber output images, initially those images are downsampled from 180×180 to 45×45 by average pooling. Then these downsampled images are flattened to 2025 features, and the ridge classification algorithm from Python's scikit-learn library is used for the determination of RWs with an L2 regularization strength, alpha, set to 3000. For all-optical classification experiments, the output beam shape is imaged onto the camera without grating dispersion, and the center of mass, (x_c, y_c), of the beam shape, $G(x, y)$, is calculated by $x_c = \frac{\sum_{x=-\infty}^{\infty} x\, G(x)}{\sum_{x=-\infty}^{\infty} G(x)}$, $y_c = \frac{\sum_{y=-\infty}^{\infty} y G(y)}{\sum_{y=-\infty}^{\infty} G(y)}$. This is the same information as the one provided by simple beam location sensors. In the first iteration, the classification line in 2D is drawn similarly with the ridge classification algorithm, but by only using two features and only training for the first iteration, keeping the classification line fixed for the rest of the iterations. For each iteration, the accuracy in the training set becomes another sampling point for the surrogate model. This model is a cubic radial basis function with a linear tail, implemented with Python Surrogate Optimization Toolbox (pySOT) and initiated by sampling $2M + 1$ Latin hypercube points, M being the number of parameters to be optimized. After the initial fixed sampling, DYCORS[55] sampling strategy explores the parameter space for the optimal set of parameters.

Disclosures

Among the authors of this article, I.O., U.T., C.M., and D.P. are listed as inventors on the patent application WO2022136146, which is still under consideration, created by École Polytechnique Fédérale de Lausanne on modifying nonlinear data transform inside waveguides. The other authors declare no competing interests.

Code and Data Availability

All relevant code and the data that support the plots within this article and other findings of this study are available from the corresponding author upon reasonable request.

References

1. B. Fasel, "An introduction to bio-inspired artificial neural network architectures," *Acta Neurol. Belg.* **103**(1), 6–12 (2003).
2. J. Kilian and H. T. Siegelmann, "The dynamic universality of sigmoidal neural networks," *Inf. Comput.* **128**(1), 48–56 (1996).
3. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature* **323**(6088), 533–536 (1986).
4. W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: scaling to trillion parameter models with simple and efficient sparsity," arXiv:2101.03961 (2022).
5. S. Reed et al., "A generalist agent," arXiv:2205.06175 (2022).
6. L. Bernstein et al., "Freely scalable and reconfigurable optical hardware for deep learning," *Sci. Rep.* **11**(1), 3144 (2021).
7. L. Ulrich, "GM bets big on batteries: a new $2.3 billion plant cranks out Ultium cells to power a future line of electric vehicles," *IEEE Spectr.* **57**(12), 26–31 (2020).
8. X. Lin et al., "All-optical machine learning using diffractive deep neural networks," *Science* **361**(6406), 1004–1008 (2018).
9. Y. Luo et al., "Design of task-specific optical systems using broadband diffractive neural networks," *Light Sci. Appl.* **8**(1), 112 (2019).
10. T. Zhou et al., "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," *Nat. Photonics* **15**(5), 367–373 (2021).
11. O. Kulce et al., "All-optical information-processing capacity of diffractive surfaces," *Light Sci. Appl.* **10**(1), 25 (2021).
12. T. Wang et al., "An optical neural network using less than 1 photon per multiplication," *Nat. Commun.* **13**(1), 123 (2022).
13. Y. Shen et al., "Deep learning with coherent nanophotonic circuits," *Nat. Photonics* **11**(7), 441–446 (2017).
14. C. Huang et al., "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits," *APL Photonics* **5**(4), 040803 (2020).
15. N. U. Dinc, D. Psaltis, and D. Brunner, "Optical neural networks: the 3D connection," *Photoniques* **104**, 34–38 (2020).
16. M. A. Al-Qadasi et al., "Scaling up silicon photonic-based accelerators: challenges and opportunities," *APL Photonics* **7**(2), 020902 (2022).
17. T. Zhou et al., "*In situ* optical backpropagation training of diffractive optical neural networks," *Photonics Res.* **8**(6), 940–953 (2020).
18. R. Shao et al., "Generalized robust training scheme using genetic algorithm for optical neural networks with imprecise components," *Photonics Res.* **10**(8), 1868–1876 (2022).
19. H. Zhang et al., "Efficient on-chip training of optical neural networks using genetic algorithm," *ACS Photonics* **8**(6), 1662–1672 (2021).
20. P. Jiang, C. A. Shoemaker, and X. Liu, "Time-varying hyperparameter strategies for radial basis function surrogate-based global optimization algorithm," in *IEEE Int. Conf. Ind. Eng. and Eng. Manage. (IEEM)*, pp. 984–988 (2017).
21. R. Turner et al., "Bayesian optimization is superior to random search for machine learning hyperparameter tuning: analysis of the blackbox optimization challenge 2020," arXiv:2104.10201 (2021).
22. D. Eriksson, D. Bindel, and C. A. Shoemaker, "pySOT and POAP: an event-driven asynchronous framework for surrogate optimization," arXiv:1908.00420 (2019).
23. Y. Li et al., "Hyper-parameter optimization using MARS surrogate for machine-learning algorithms," *IEEE Trans. Emerg. Top. Comput. Intell.* **4**(3), 287–297 (2020).
24. C. Cartis, L. Roberts, and O. Sheridan-Methven, "Escaping local minima with derivative-free methods: a numerical investigation," arXiv:1812.11343 (2019).
25. G. Tanaka et al., "Recent advances in physical reservoir computing: a review," *Neural Netw.* **115**, 100–123 (2019).
26. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing* **70**, 489–501 (2006).
27. G. C. Valley et al., "Photonic reservoir computer using speckle in multimode waveguide ring resonators," *Opt. Express* **29**(13), 19262–19277 (2021).
28. G. C. Valley et al., "Classification of time-domain waveforms using a speckle-based optical reservoir computer," *Opt. Express* **28**(2), 1225–1237 (2020).

29. M. Rafayelyan et al., "Large-scale optical reservoir computing for spatiotemporal chaotic systems prediction," *Phys. Rev. X* **10**(4), 041037 (2020).

30. D. Brunner et al., "Parallel photonic information processing at gigabyte per second data rates using transient states," *Nat. Commun.* **4**(1), 1364 (2013).

31. F. Duport et al., "All-optical reservoir computing," *Opt. Express* **20**(20), 22783 (2012).

32. A. Dejonckheere et al., "All-optical reservoir computer based on saturation of absorption," *Opt. Express* **22**(9), 10868–10881 (2014).

33. L. G. Wright et al., "Deep physical neural networks trained with backpropagation," *Nature* **601**(7894), 549–555 (2022).

34. J. Pauwels et al., "Distributed Kerr non-linearity in a coherent all-optical fiber-ring reservoir computer," *Front. Phys.* **7**, 138 (2019).

35. U. Teğin et al., "Scalable optical learning operator," *Nat. Comput. Sci.* **1**(8), 542–549 (2021).

36. O. Tzang et al., "Adaptive wavefront shaping for controlling nonlinear multimode interactions in optical fibres," *Nat. Photonics* **12**, 368–374 (2018).

37. E. Deliancourt et al., "Wavefront shaping for optimized many-mode Kerr beam self-cleaning in graded-index multimode fiber," *Opt. Express* **27**(12), 17311–17321 (2019).

38. U. Teğin et al., "Controlling spatiotemporal nonlinearities in multimode fibers with deep neural networks," *APL Photonics* **5**(3), 030804 (2020).

39. J. Cao, Z. Lin, and G.-B. Huang, "Self-adaptive evolutionary extreme learning machine," *Neural Process. Lett.* **36**(3), 285–305 (2012).

40. Z. Liu et al., "Deep learning face attributes in the wild," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 3730–3738 (2015).

41. A. Mafi, "Pulse propagation in a short nonlinear graded-index multimode optical fiber," *J. Lightwave Technol.* **30**(17), 2803–2811 (2012).

42. U. Teğin et al., "Reusability report: predicting spatiotemporal nonlinear dynamics in multimode fibre optics with a recurrent neural network," *Nat. Mach. Intell.* **3**(5), 387–391 (2021).

43. A. Klein et al., "Meta-surrogate benchmarking for hyperparameter optimization," arXiv:1905.12982 (2019).

44. H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," arXiv:1708.07747 (2017).

45. Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.* **1**(4), 541–551 (1989).

46. T. Rahman et al., "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Comput. Biol. Med.* **132**, 104319 (2021).

47. M. Tan and Q. E. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," arXiv:1905.11946 (2020).

48. V. Rotemberg et al., "A patient-centric dataset of images and meta-data for identifying melanomas using clinical context," *Sci. Data* **8**(1), 34 (2021).

49. ISIC 2020 Challenge Dataset, https://challenge2020.isic-archive.com/ (accessed 12 June 2022).

50. D. Psaltis et al., "Accurate numerical computation by optical convolution," *Proc. SPIE* **0232**, 151–156 (1980).

51. P. Kora et al., "Transfer learning techniques for medical image analysis: a review," *Biocybern. Biomed. Eng.* **42**(1), 79–107 (2022).

52. Z. Liu et al., "Kerr self-cleaning of femtosecond-pulsed beams in graded-index multimode fiber," *Opt. Lett.* **41**(16), 3675–3678 (2016).

53. "DLP7000 data sheet, product information and support TI.com," https://www.ti.com/product/DLP7000 (accessed 14 July 2022).

54. C. Yao et al., "Evaluating and analyzing the energy efficiency of CNN inference on high-performance GPU," *Concurr. Comput. Pract. Exp.* **33**(6), e6064 (2021).

55. R. G. Regis and C. A. Shoemaker, "Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization," *Eng. Optim.* **45**(5), 529–555 (2013).

Ilker Oguz is a doctoral student at the Doctoral Program in Photonics, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Currently, he works on efficient physical computing architectures and training algorithms. Previously, he finished his bachelor's degree in electrical engineering at Middle East Technical University (METU), Turkey, in 2018, and his master's degree in bioimaging in the Department of Information Technology and Electrical Engineering, ETH Zürich, Switzerland, in 2020, as an awardee of Excellence Scholarship and Opportunity Program.

Jih-Liang Hsieh is a PhD student at Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland. He specializes in developing photonics neural networks and optical computing systems. He employs both nonlinear and linear optical manipulation techniques to harness the power of light in either optical fibers or free space. His work aims to revolutionize computing paradigms and drive innovation toward a more sustainable and energy-efficient future for computational technologies.

Niyazi Ulas Dinc is a postdoctoral researcher at École Polytechnique Fédérale de Lausanne, Switzerland. He obtained his BSc degrees in electrical engineering and physics from METU, Turkey, his MSc and PhD degrees in microengineering from EPFL, Switzerland. He is currently working on computer-generated optical volume elements to achieve the desired mapping of arbitrary input–output fields in the spatiotemporal domain for computing and beam shaping purposes.

Uğur Teğin is an assistant professor in electrical and electronics engineering at Koç University. He received his BSc and MSc degrees from Bilkent University, Ankara, Turkey, in 2015 and 2018, respectively. He obtained his PhD from EPFL, Lausanne, Switzerland, in 2021. He then pursued his postdoctoral studies in medical and electrical engineering at the California Institute of Technology, United States. His research interests include nonlinear optics, optical computing, machine learning, fiber optics lasers, and ultrafast optics.

Mustafa Yildirim earned his bachelor's degree from the Middle East Technical University and followed up with his master's degree from EPFL, both in electrical engineering. Currently, he is engaged as a doctoral candidate at EPFL's Doctoral Program in Photonics. His primary focus lies in advancing optics-based neural network architectures, aiming for efficient and sustainable compute solutions.

Carlo Gigli is postdoctoral researcher at the Laboratory of Applied Photonics Devices, EPFL, Lausanne. He received his MS degree in physical engineering from Politecnico di Torino and Université Paris Diderot in 2017, and his PhD in physics from Université de Paris with a thesis on the design, fabrication, and characterization of dielectric resonators and metasurfaces for nonlinear optics. His current research activity focuses on AI-assisted bioimaging and photonic devices design.

Christophe Moser is a full professor at the Institute of Electrical and Microengineering and currently the director of the Microengineering Section at EPFL. He obtained his physics diploma degree from EPFL and his PhD from California Institute of Technology. He was the CEO of Ondax, Inc., prior to joining EPFL. His current research topics include light-based additive manufacturing—tomographic volumetric, two photon—and neuromorphic computing using linear and nonlinear propagation in optical fibers.

Demetri Psaltis received his BSc, MSc, and PhD degrees from Carnegie-Mellon University, Pittsburgh, Pennsylvania, United States. He is a professor of optics and a director of the Optics Laboratory at EPFL, Switzerland. His research interests include imaging, holography, biophotonics, nonlinear optics, and optofluidics. He has authored or co-authored more than 400 publications in these areas. He was the recipient of the International Commission of Optics Prize, the Humboldt Award, the Leith Medal, and the Gabor Prize.