



Physics-aware Differentiable Discrete Codesign for Diffractive Optical Neural Networks

Yingjie Li, Ruiyang Chen, Weilu Gao, Cunxi Yu

ECE Department, University of Utah, Salt Lake City, Utah, 84112, USA

{yingjie.li, cunxi.yu}@utah.edu

ABSTRACT

Diffractive optical neural networks (DONNs) have attracted lots of attention as they bring significant advantages in terms of power efficiency, parallelism, and computational speed compared with conventional deep neural networks (DNNs), which have intrinsic limitations when implemented on digital platforms. However, inversely mapping algorithm-trained physical model parameters onto real-world optical devices with discrete values is a non-trivial task as existing optical devices have non-unified discrete levels and non-monotonic properties. This work proposes a novel device-to-system hardware-software codesign framework, which enables efficient physics-aware training of DONNs w.r.t arbitrary experimental measured optical devices across layers. Specifically, Gumbel-Softmax is employed to enable differentiable discrete mapping from real-world device parameters into the forward function of DONNs, where the physical parameters in DONNs can be trained by simply minimizing the loss function of the ML task. The results have demonstrated that our proposed framework offers significant advantages over conventional quantization-based methods, especially with low-precision optical devices. Finally, the proposed algorithm is fully verified with physical experimental optical systems in low-precision settings.

KEYWORDS

Optical neural network, diffractive optical neural networks, hardware-software co-design, Gumbel-Softmax

ACM Reference Format:

Yingjie Li, Ruiyang Chen, Weilu Gao, Cunxi Yu. 2022. Physics-aware Differentiable Discrete Codesign for Diffractive Optical Neural Networks. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD '22)*, October 30-November 3, 2022, San Diego, CA, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3508352.3549378>

1 INTRODUCTION

During the past half-decade, there has been significant growth in machine learning with deep neural networks (DNNs). DNNs improve productivity in many domains such as large-scale computer vision, natural language processing, and data mining tasks [24, 34, 37]. However, conventional DNNs implemented on digital platforms have intrinsic limitations in computation and memory requirements [1, 20, 35]. Moreover, when it deals with computation-intensive tasks, its energy cost will be a great concern. To overcome

limitations in resources and find an energy-saving computation method, people have turned their eyes to optics [11–14, 25, 26, 28, 43, 44]. Specifically, the free-space diffractive optical neural networks (DONNs), which is based on light diffraction, featuring millions of neurons in each layer interconnected with neurons in neighboring layers, show its great potential in improving efficiency in computing with neural networks [28]. More importantly, Chen et al. [4], Li et al. [26], Li and Yu [27], Lin et al. [28], Rahman et al. [33], Shen et al. [36] demonstrated that diffractive models controlled by physical parameters are differentiable, such that the parameters can be optimized with conventional automatic differentiation engines.

However, when the DONN system is deployed on physical hardware, it shows significant accuracy degradation compared to the numerical physics emulation [28, 44], e.g., the accuracy degradation is claimed as 30% in [44]. To narrow the algorithm-hardware mismatch gaps between differentiable numerical physics models and physical systems, hardware-software codesign training algorithms are needed to deal with the practical response of optical devices. For example, the reconfigurability of DONNs is implemented using spatial light modulators (SLMs), which have a discrete and non-monotonic complex-valued modulation of propagating optical fields as a function of applied voltages with finite-precision [44].

Therefore, despite the diffraction propagation in the DONN system is differentiable, directly adding discrete mapping from device to DONN systems will break the gradient chain in backpropagation. Moreover, in optical hardware systems, diffractive layers implemented with analog optical devices can behave differently due to different optical configurations or device responses, i.e., non-uniformity exists across the compute units (devices), while the DONN model is trained and optimized on digital platforms with uniform and stable number represented computation. Thus, to narrow the gap between numerical emulation and practical deployment, while training a multi-layer DONNs system, there is a great need to develop a flexible training framework that can optimize the DONNs parameters w.r.t various optical devices from layer to layer. While quantization techniques are applicable to discrete mapping from the device level to DONNs system, there are several critical limitations due to the fact that optical devices used in DONNs are analog, non-monotonic, and non-unified. Specifically, the trainable parameters in DONN systems are not only limited to discrete mappings with irregular and analog device responses but more importantly limited to the constraints in physics. For example, the phase for the light wave is a periodic function with 2π as the period. Thus, the trainable parameters w.r.t the phase modulation devices in the DONN system should be restricted within $[0, 2\pi]$ and aware of the 2π -period during the training process [17, 28, 44].

This work studies an efficient and flexible codesign framework that enables physics-aware differentiable discrete mappings from



devices to DONN systems via Gumbel-Softmax (GS) [15, 19, 29], which has been recently experimentally verified on a visible DONNs hardware platform [3]. This approach can overcome the aforementioned limitations with GS enabling fully differentiable discrete mapping regardless of the number representation format, range, or discrete distribution. Specifically, our results demonstrate the advantages over existing state-of-the-art DONNs training approaches and quantization algorithms using real-world optical devices, in various DONNs architecture settings. Moreover, we perform comprehensive temperature scheduling exploration and statistical analysis in the GS algorithm to offer insights of this framework. Our results demonstrate the substantial advantages for DONNs co-design in image classification, particularly when deployed optical devices are limited to low precision. Finally, we verify the proposed approach in our visible range DONNs hardware platform [3] in low precision settings.

2 BACKGROUND

Diffraction Optical Neural Networks (DONNs) – Recently, there have been increasing efforts on optical neural networks and optics-based DNNs hardware accelerators, which bring significant advantages for machine learning systems in terms of their power efficiency, parallelism and computational speed, demonstrated at various optical computing systems [6, 9, 10, 26, 28, 31, 33, 36, 38]. Among them, free-space *diffraction optical neural networks* (DONNs), which is based on the light diffraction, features millions of neurons in each layer interconnected with neurons in neighboring layers. The ultrahigh density and parallelism make this system possess fast and high throughput computing capability. One of the significant advantages of DONNs is the computational density and energy efficiency, where such a platform can be scaled up to millions of artificial neurons while with much less energy cost compared to conventional DNNs on digital platforms [26, 28, 30, 31].

In conventional DNNs, forward propagations are computed by generating the feature representation with floating-point weights associated with each neural layer. While in DONNs, such floating-point weights are encoded in the **complex-valued transmission coefficient of each neuron** in diffractive layers and **free-space propagation function** [5], which is multiplied onto the light wavefunction as it propagates through the neuron to next diffractive layer. Specifically, in the numerical emulation for DONN systems, there are two sets of network parameters. One is for diffraction approximation, which is non-trainable parameters describing the propagation after the light wave is diffracted at the diffractive layers, and defined by the natural physics phenomenon. The diffraction propagation connects neurons between layers. The other one is for phase modulation, which is our targeted “trainable parameters” in DONNs. By applying phase modulation to the input light wave at each diffractive layer, the distribution of the light intensity at the end of the system will be modified accordingly.

Similar to conventional DNNs, the final output class of DONN systems is predicted based on generating labels according to a given one-hot representation, i.e., the max energy reading over the output light intensity of the last layer observed by detectors, where the loss function is at the same time calculated. Thus, the DONN system can be trained by optimizing parameters for phase modulation in

each diffractive layer w.r.t specific machine learning (ML) task loss functions, e.g., image classifications. Details of DONNs training and inference are provided in Section 4.2.

However, when such DONNs systems are deployed on physical hardware, existing training approaches do not take real device response (see Figure 1c) into consideration and only assume simple phase-only modulation without any limitations. Ideal assumption and quantization errors create miscorrelation gaps between numerical models and hardware deployment, leading to significant accuracy degradation such as 30% drop on MNIST dataset [44].

Physical optical devices for phase modulation – The DONN systems will be deployed on physical hardware with analog optical devices after trained on digital platforms. Optical devices functioning as diffractive layers are expected to provide light diffraction and around 2π phase modulation over full function range in the DONN system. For example, the Spatial Light Modulators (SLMs)¹ made with twisted nematic (TN) liquid crystal can provide a phase shift of about 2π at 450nm, about 1.8π at 532nm and around 1π at 800nm w.r.t the full range (e.g., 256 discrete voltage stages) of supplying control voltage. For systems with laser wavelength in Terahertz (THz) range, i.e., 0.1mm to 1mm, a 3D printed mask with designed thickness at each pixel made with UV-curable resin can be used as the diffractive layers in THz optical systems [28].

For example, in our DONN system [3], the diffractive layers are implemented with SLMs. Specifically, the SLM is an array of twisted nematic (TN) liquid crystals, which can be twisted to different angles by different applied control voltages, providing different phase modulation for the input light beam. Each pixel in the SLM is a liquid crystal independently controlled by the control voltage, which can be customized by users via HDMI. A practical optical response in amplitude and phase modulation of the SLM with eight discrete control voltages is provided in Figure 1c, and we can see that the optical responses w.r.t discrete control voltages is not unified over its functioning range. Additionally, each single SLM can response differently even under identical experimental setups due to fabrication errors. When multiple SLMs are employed in one system, the error will accumulate, and worsening the correlation between the numerical emulations and the physical hardware experiments, which highlights the importance and motivation of our proposed hardware-software codesign framework.

Gumbel-Softmax – Gumbel-Softmax is a continuous distribution on the simplex which can be used to approximate discrete samples [15, 19, 29]. With Gumbel-Softmax, discrete samples can be differentiable and their parameter gradients can be easily computed with standard backpropagation. Let z be the discrete sample with one-hot representation with k dimensions and its class probabilities are defined as $\pi_1, \pi_2, \dots, \pi_k$. Then, according to the Gumbel-Max trick proposed by Gumbel [15], the discrete sample z can be presented by:

$$z = \text{one_hot}(\arg\max_i [g_i + \log \pi_i]) \quad (1)$$

where g_i are i.i.d samples drawn from Gumbel(0, 1). Then, we can use the differentiable approximation Softmax to approximate the one-hot representation for z , i.e., $\nabla_{\pi} z \approx \nabla_{\pi} y$:

$$y_i = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{i=1}^k \exp((\log(\pi_i) + g_i)/\tau)} \quad (2)$$

¹<https://holoe.com/lc-2012-spatial-light-modulator>

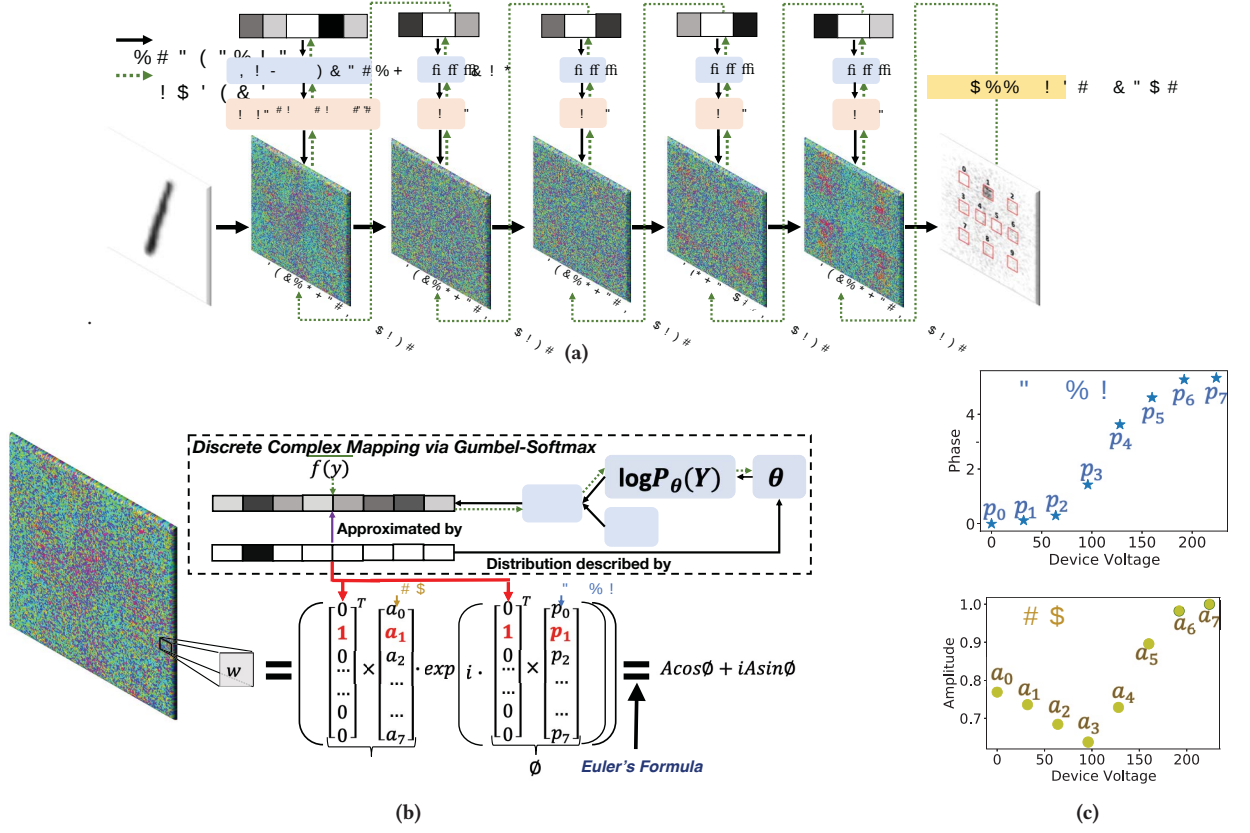


Figure 1: Illustration of the proposed GS framework in training a 5-layer DONN system – (a) During inference, input signal is modulated by each layer in both amplitude and phase encoded by optical devices controlled by discrete voltage values. In backpropagation for training, the gradients will flow through Gumbel-Softmax over the approximated differentiable representation of the one-hot encoded discrete voltage levels; (b) Detailed illustration of discrete device-to-system mapping via Gumbel-Softmax in both forward and backpropagation; (c) Illustration of real-world device responses.

where $i = 1, 2, \dots, k$. The softmax temperature τ is introduced to modify the distributions over discrete levels. Softmax distributions will become more discrete and identical to one-hot encoded discrete distribution as $\tau \rightarrow 0$, while at higher temperatures, the distribution becomes more uniform as $\tau \rightarrow \infty$ [19]. Gumbel-Softmax distributions have a well-defined gradient $\frac{\partial y}{\partial \pi}$ w.r.t the class probability π . When we replace discrete levels with Gumbel-Softmax distribution depending on its class probability, we are able to use backpropagation to compute gradients. Recently, Gumbel-softmax has been applied to differentiable neural architecture search [7, 8, 18, 40, 41] and differentiable quantization [2]. However, enabling differentiable discrete weight training in a fully physics-differentiable neural networks with real-world physical system has not yet been studied. Particularly, the parameters in DONNs are limited to non-negative values due the nature of optical physics.

3 APPROACH

System Overview and GS-based Training Algorithm – We illustrate the proposed training framework using a five-layer DONN system implemented for a ten-class image classification task, with ten detector regions placed evenly on the detector plane. As shown

in Figure 1, each diffractive layer modifies the amplitude and phase of the input light signal. In our experimental setup, diffractive layers are implemented with SLMs. As we discussed in Section 2, the analog optical device SLMs provide modulation w.r.t the discrete input control voltage values, and the physical responses for each SLM can be different and non-unified, which requires a hardware-software codelign algorithm for precise emulation.

To enable differentiable discrete mapping, *Gumbel-Softmax* is added in the numerical modeling of DONNs. First, our framework defines the input discrete voltage values as trainable parameters, where each pixel is represented using a one-hot vector. The trainable parameters dimensions are then defined by (1) the system size and (2) the number of discrete values in the devices. For example, let the system size be 200×200 with 8 discrete levels in the devices, the trainable voltage parameters will be $200 \times 200 \times 8$ in each layer. Note that the phase and amplitude modulation will still be in the shape of 200×200 , where the optical properties are mapped by matmul the one-hot vectors ($200 \times 200 \times 8$) and the device level vector (8×1). Then, to deal with the problem of gradient chain breakage brought by the discrete trainable parameters, in Gumbel-Softmax, a Gumbel distribution $g \sim \text{Gumbel}(0, 1)$ and a class probability θ for

the discrete levels are introduced to approximate the discrete levels. As it is shown in Figure 1b, during backpropagation in the training process, instead of propagating gradients to the discrete one-hot voltage levels directly, it will propagate through the differentiable approximation to the discrete levels generated by Gumbel-Softmax distribution with its class probability θ . The differentiable approximation will be updated according to the training algorithm, and the discrete voltage levels will be updated by its class probability θ from approximation. The process in Figure 1b can be described as:

$$w^{i,j} = \text{one_hot}\left(\frac{\exp((\log(\theta_n) + g_n)/\tau)}{\sum_{n=1}^k \exp((\log(\theta_n) + g_n)/\tau)}\right), \quad (3)$$

$$g_n \sim \text{Gumbel}(0, 1), \quad k = \text{total discrete levels}, \quad i, j \in [0, N - 1]$$

where $w^{i,j}$ be the voltage level applied to the pixel located at $[i, j]$ in the diffractive layer with a size $N \times N$. In Equation 3, one_hot will pick the class with the highest probability over k discrete voltage levels after softmax as 1, while other $k - 1$ stages will be 0 in the one-hot representation for $w^{i,j}$.

$$w_{\mathbb{C}}^{i,j} = \underbrace{(w^{i,j} \cdot A)}_{\text{Matmul}} \cdot \underbrace{\cos(w^{i,j} \cdot P)}_{\text{Real}} + i \underbrace{(w^{i,j} \cdot A)}_{\text{Matmul}} \cdot \underbrace{\sin(w^{i,j} \cdot P)}_{\text{Imaginary}}, \quad (4)$$

for $i, j \in [0, N - 1]$

Let A be the array with discrete calibrated amplitude value, P be the array with discrete calibrated phase value. $w_{\mathbb{C}}^{i,j}$ will be the phase and amplitude modulation provided by the pixel located at $[i, j]$, which is a complex number resulting from mapping the one hot represented voltage level with the amplitude and phase responses of the specific SLM (e.g., the optical responses in Figure 1c). Let $w^{i,j} \in W$, $w_{\mathbb{C}}^{i,j} \in W_{\mathbb{C}}$, where $i, j \in [0, N - 1]$, N is the size of diffractive layer. According to Equation 3, the discrete variable W has the distribution depending on θ and forward function $f(W)$. The objective is to minimize the expected cost $L(\theta) = \mathbb{E}_{W \sim p_{\theta}(W)} [f(W)]$, which is the ML loss in the system, e.g., in DONNs system for image classification, L is usually set as the MSE Loss [28], [44], via gradient descent, which requires us to estimate $\nabla_{\theta} \mathbb{E}_{W \sim p_{\theta}(W)} [f(W)]$. The discrete sample W can be approximated by $G(\theta, g)$. The gradients from f to θ will be computed as follows:

$$\begin{aligned} \frac{\partial}{\partial \theta} \mathbb{E}_{W \sim p_{\theta}(W)} [f(W)] &= \frac{\partial}{\partial \theta} \mathbb{E}_g [f(G(\theta, g))] \\ &= \mathbb{E}_{g \sim \text{Gumbel}(0,1)} \left[\frac{\partial f}{\partial G} \frac{\partial G}{\partial \theta} \right] \end{aligned} \quad (5)$$

To be more specific, as shown in Figure 1b, for each pixel ($w_{\mathbb{C}}^{i,j}$) in diffractive layers, the modulation provided by the pixel is represented by complex tensors that emulate the diffraction of light. The complex number is transformed from phase and amplitude modulation by *Euler's formula*. Since the input light signal is also described by complex numbers, the modulation can be easily realized by multiply the two complex numbers.

As a results, after the input image is encoded with a coherent laser beam in complex domain, where the imaginary part is initialized as all zeros, the emulations/forward function of the DONN system for image classification can be described as follows – Each

diffractive layer is composed with dense diffraction units with apertures and phase modulators. Specifically, the aperture will provide **light diffraction**, which connects the neurons in neighboring diffractive layers and the diffraction is non-trainable parameters w.r.t user-defined setups in the system. The **phase modulation** embedded in each aperture functions as 'neurons' in DONN to modify the input light wave, and the phase modulation is our targeted trainable parameters in the system. In the emulation for the physics-based DONN system, we employ (1) Fresnel approximation [5], which is a mathematical approximation for light diffraction through an aperture, in the forward function to describe the light diffraction in DONN systems; (2) complex-valued matrix multiplication for modulator parameters and the input light wavefunction to describe the phase modulation in the system.

For light diffraction, the input at point (x, y) at l -th diffractive layer can be seen as the summation of the outputs at $(l - 1)$ -th layer over the plane (x', y') , i.e.,

$$f_l^1(x, y, z) = \iint f_{l-1}^0(x', y', 0) h(x - x', y - y', z) dx' dy' \quad (6)$$

where z is the distance between layers, h is the impulse response function of free space. f_{l-1}^0 is the output wavefunction of points on $(l - 1)$ -th layer and also the input for free-space propagation, f_l^1 is the output function of the free-space propagation and the input to the phase modulation at l -th layer. Equation 6 can be calculated with spectral algorithm, where we employ Fast Fourier Transform (FFT) for fast and differentiable computation. By convolution theorem, the integral over x, y -axes of the convolution of f_{l-1}^0 and h is the product of 2D Fourier transformations over x, y -axes of f_{l-1}^0 and h , i.e.,

$$\mathcal{F}_{xy}(f_l^1(x, y, z)) = \mathcal{F}_{xy}(f_{l-1}^0(x', y', 0)) \mathcal{F}_{xy}(h(x, y, z)) \quad (7)$$

$$U_l(\alpha, \beta, z) = U_{l-1}(\alpha, \beta, z) H(\alpha, \beta, z) \quad (8)$$

where U and H are the Fourier transformation of f and h respectively. The impulse function used for Fresnel approximation is

$$h(x, y, z) = \frac{\exp(ikz)}{i\lambda z} \exp\left\{\frac{ik}{2z}(x^2 + y^2)\right\} \quad (9)$$

where $i = \sqrt{-1}$, λ is the wavelength of the laser source, $k = 2\pi/\lambda$ is free-space wavenumber.

For phase modulation, we model the output of the free-space propagation at l -th layer from Equation 8 as $U_l(\alpha, \beta, z)$, which is then converted back to spatial domain (x, y) via inverse FFT as $f_l^2(x, y, z)$ as the input to the phase modulation at l -th layer. The phase modulation is described in Equation 4 as $W_{\mathbb{C}}$. As we discussed in Section 2, the phase modulation provided by each diffraction unit in the layer is independently configured, thus, the phase modulation $W_{\mathbb{C}}$ functions w.r.t the location (x, y) at the l -th layer. Thus, the output wavefunction after phase modulation is expressed as

$$f_{l+1}^0(x, y, 0) = f_l^2(x, y, z) \times W_{\mathbb{C}}^l(x, y) \quad (10)$$

where $f_{l+1}^0(x, y, 0)$ is the input wavefunction for the forward function for $(l + 1)$ -th layer.

As a result, we model the light diffraction (Equations 6–9), phase modulation (Equation 10), and device-to-system codesign (Equations 3–4) in fully differentiable numerical formats. Thus, the DONNs can be trained with conventional *autograd* algorithms by simply

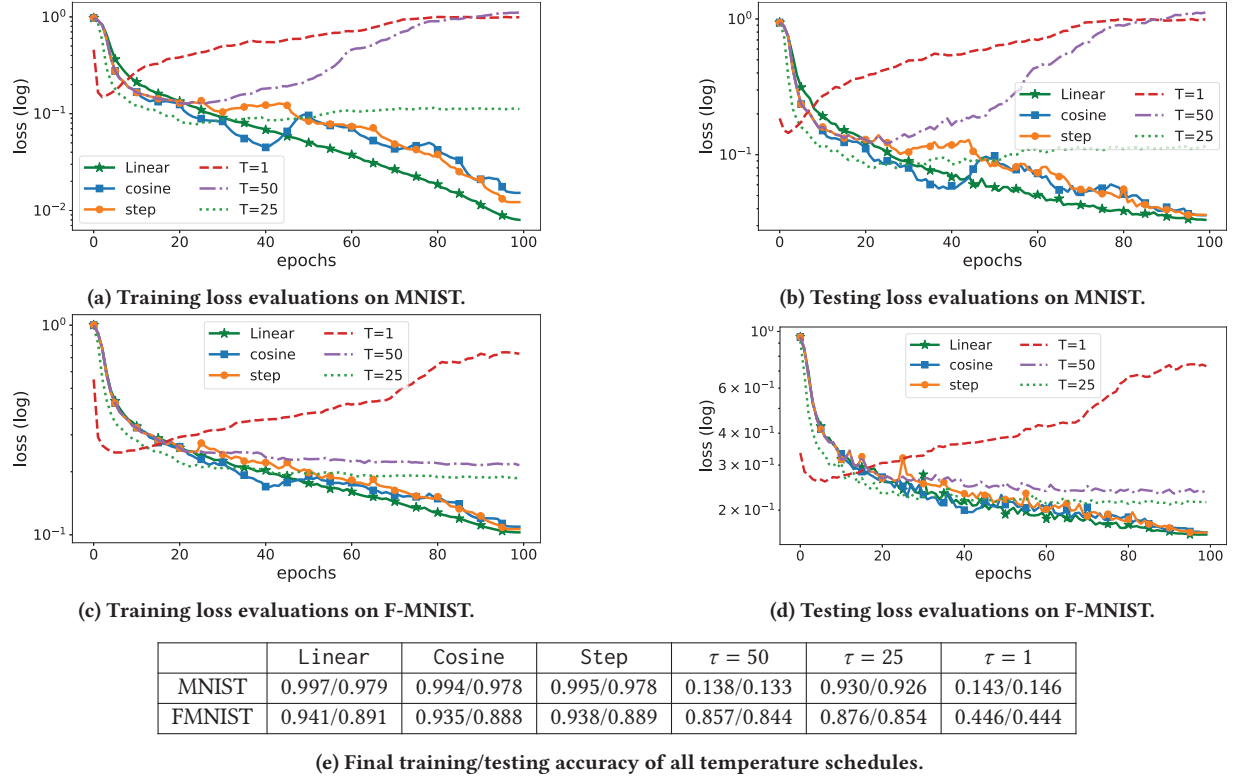


Figure 2: Evaluations of temperature scheduling in GS-based discrete device-to-system DONNs training framework.

minimizing DONNs ML loss function, e.g., Adam algorithm [21]. Note that in the DONN system, since diffractive layers are propagated in sequence, each layer can be implemented using different devices, i.e., with different calibrated data points. As we can see in Figure 1, mapping multiple devices in one DONN system can be simply realized using the proposed framework by replacing the *Amp* and *Phase* vectors at each layer.

Gumbel-Softmax Exploration – As it is discussed in Section 2, the variance of the approximated Gumbel-Softmax distribution over discrete levels is determined by τ , which is also referred as *temperature* in Gumbel-Softmax. While giving a higher temperature value, Gumbel-Softmax will result in less variance that is close to a uniform distribution. In opposite, when τ is close to 0, it will be more variant over discrete levels, i.e., be more identical to one-hot distribution. Thus, similar to simulated annealing, the algorithm should be first deployed with high temperature to enable coarse-grain global search. The temperature should then be annealed down to shrink the search space to find the local optimized point. Specifically, at the early training stage, we expect the variance between different levels to be small, such that the discrete values are easier to be changed during gradient descent optimization. As the optimization efforts increase, it is expected to decrease the temperature to fine-tune the optimization, where most of the discrete values are far more stable during gradient descent optimization. In Section 4.1, we explore and provide comprehensive discussions on six different temperature schedules.

4 RESULTS

System Parameters – The default system used in this work is designed with five diffractive layers with the size of 200×200 , i.e., the size of layers and the size of total ten detector regions are 200×200 . To fit the optical system, the original input images from MNIST [23] and FashionMNIST (FMNIST) [42] with size of 28×28 will be interpolated into size of 200×200 and encoded with the laser source whose wavelength is 532 nm . The physical distances between layers, first layer to source, and final layer to detector, are set to be 27.94 cm . As shown in Figure 1, ten separate detector regions for ten classes are placed evenly on the detector plane with the size of 20×20 , where the sums of the intensity of these ten regions are equivalent to a 1×10 vector in float32 type. The final prediction results will be generated using `argmax`. The default DONN system is implemented with an optical device consisting of 8 discrete voltage values for modulation. **Training Setups** – The learning rate in the training process is 0.5 trained with 100 epochs for all experiments using Adam [21] with batch size 500. The implementations are constructed using PyTorch v1.8.1. All experimental results are conducted on an Nvidia 2080 Ti GPU.

4.1 Temperature scheduling in Gumbel-Softmax

While deploying Gumbel-Softmax to enable differentiable discrete training for DONNs, the temperature value in Gumbel-Softmax is known as an important hyperparameter for the training performance. Thus, we first explore different temperature schedules in the proposed training framework for implementing DONNs with

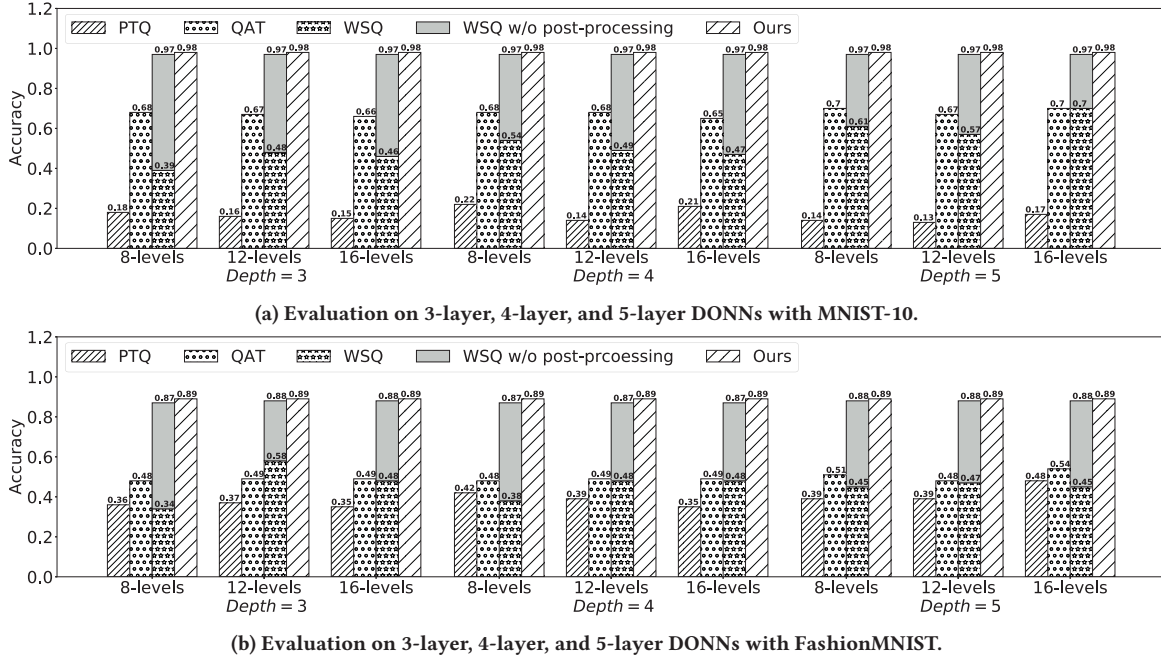


Figure 3: Accuracy performance comparisons between the proposed framework and conventional quantization methods, i.e., PTQ [22], QAT [22], and WSQ [16, 39], for discrete device mapping in 3-, 4-, and 5-layer DONNs using MNIST and FMNIST datasets. WSQ w/o post-processing shows the accuracy performance of the model quantized with both negative and positive discrete levels. Ours denotes the performance of the model quantized with the proposed GS framework.

low-precision optical devices. Specifically, the results shown in Figure 2 are trained with experimentally measured real-world devices with 8 discrete values (see Figure 1c).

As discussed in Section 3, we mimic the concept of temperature scheduling in *simulated annealing* into the Gumbel-Softmax based training framework. We evaluate six different temperature schedules, which all start with highest temperature $\tau_h = 50$, and end with lowest temperature $\tau_l = 1$ with 100 training epochs (Figure 2). First, we set three static temperature training as baselines for comparisons, which are trained with static temperatures for the whole training process, i.e., (1) $\tau = 1$, (2) $\tau = 25$, and (3) $\tau = 50$. For dynamic temperature scheduling, we evaluate (4) linear temperature decaying scheduling (Linear), with temperature decay rate as 0.5 per epoch; and (5) cosine-annealing-decaying (Cosine) temperature schedule, where we set $\tau_{cosine} = [50, 40, 30, 20, 40, 30, 20, 30, 15, 5, 10, 1]$. With higher temperature, i.e., larger exploration space for the algorithm, it is expected to train more epochs. Thus, we set the training epochs for each temperature as $[10, 10, 10, 10, 10, 10, 10, 8, 7, 5, 5, 5]$; and (6) step temperature decaying (Step) the temperature schedule is set as $\tau_{step} = [50, 40, 30, 20, 10, 5, 1]$ with the training epochs per temperature as $[25, 20, 20, 15, 10, 5, 5]$.

In Figures 2(a)–(d), the training and testing processes are fully recorded for MNIST and FMNIST, with the x-axis representing the training iteration and the y-axis representing the loss value. The training and testing accuracy is shown in Figure 2(e). As we can see, for both MNIST and FMNIST, the system performs better with annealing temperature schedules in Gumbel-Softmax compared to the static temperature setups. From Figure 2, we empirically conclude

that the Linear annealed temperature schedule works best for our system, as training and testing loss converge simultaneously and most efficiently. Therefore, we deploy Linear temperature decaying in the rest experiments.

4.2 Comparisons with Quantization Methods

To demonstrate the advantages of the proposed GS-based training approach over existing methods, we compare the accuracy performance with existing approaches used in experimental optical studies [28, 44] and various quantization methods. The targeted optical device is the same 8-level device. Note that the proposed GS-based training framework is fully aware of the discrete distribution on-the-fly in the training process, in which the weights are ready for hardware deployment directly. The trained discrete values are used to configure the device, which produces the complex-valued property of optical devices.

The baseline results are generated with various quantization methods under the same optical system setup, including post-training quantization (PTQ) [22], quantization-aware training (QAT) [22], and the weights sharing quantization (WSQ) [16, 39] shown in Figure 3. For all the quantization methods compared, we first fit a multi-polynomial regression model that takes supply voltage as input and produces phase and amplitude value. The discrete values are selected w.r.t specific optical device responses. For PTQ, we train the DONNs while considering the device voltage values being feasible in float32 precision and round the voltages to the nearest available discrete points (8, 12, or 16 discrete integers) after entire training process. For QAT, instead of quantization after all training iterations, the values are rounded to the nearest available

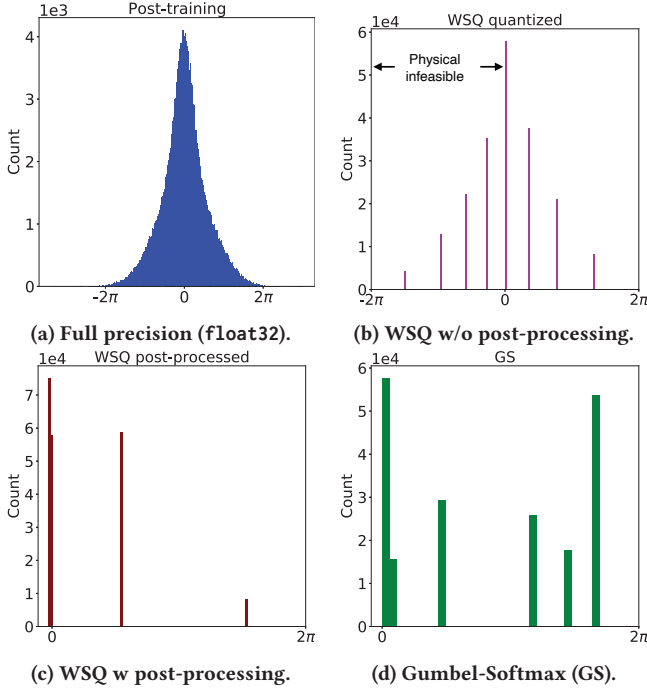


Figure 4: Weights distributions for the same 5-layer DONN model quantized with different quantization methods.

discrete values during the training process (clipping over minibatch updates), and the loss is calculated with the clipped SLMs voltages. Specifically, during the training of the entire dataset (one epoch), we explore various options for minibatch clipping, i.e., for the MNIST training dataset with 60,000 samples, we do the quantization every 12,000 samples, 15,000 samples, 20,000 samples, 30,000 samples, and 60,000 samples, respectively. Our experimental results demonstrate that QAT training set up with one quantization per epoch over the entire dataset performs the best, which is included in Figure 3. For WSQ, the input model is a pre-trained DONNs in `float32` precision, and is then quantized with weights sharing methods conducted on KMeans clustering [16, 32, 39]. Note that the quantized clusters are automatically selected by the algorithm without full awareness of the physical device. Thus, unlike GS method, which is fully aware of the physical feasibility of the devices, WSQ method requires further post-processing to exactly match the discrete levels available on the physical devices. We provide additional discussion on the physical feasibility of devices during quantization with weight distribution analysis shown in Figure 4.

As shown in Figure 3, four different approaches are evaluated with 8, 12, and 16 discrete levels in 3-, 4-, 5-layer DONNs systems, respectively. Note that the discrete values in optical devices are all non-negative numbers. Thus, models trained and quantized with WSQ algorithm require post-processing to round all negative parameters to the nearest non-negative available value. To make comprehensive comparison, we provide both quantization results before and after post-processing. We can see that models trained with GS method always perform at its best accuracy (0.98 for MNIST, 0.89 for FMNIST). Specifically, (1) within the same depth of DONN systems, the model trained with the proposed GS-based framework

demonstrates better accuracy in all cases. More importantly, for the system implemented with different discrete levels, we can see that the proposed framework is able to train the DONNs to match the best accuracy regardless of its complexity. (2) For systems with different structural complexity, GS method shows significant advantages in training, especially devices with fewer (e.g., 8-level) discrete levels. (3) For compared quantization methods, even though they work decently before post-processing, the performance will significantly degrade once post-processing is applied. For example, before post-processing, the models quantized with WSQ method perform similarly as GS method (e.g., 0.97 for MNIST, 0.88 for FMNIST). However, the accuracy after post-processing degrades more than 27% consistently for all setups. Our results have demonstrated that the proposed differentiable discrete training framework with Gumbel-Softmax offers significant training performance improvements over conventional quantization algorithms, especially in co-designing DONN systems built with optical devices with very limited and non-uniform optical responses.

To understand the performance degradation by post-processing, we explore the distribution of weight parameters (Figure 4) of 5-layer DONNs with 8-level devices. We can find that the performance gap between models before and after post-processing shown in Figure 3 comes from the unawareness of the physical optical devices, which only provides feasible positive discrete levels. Figure 4a shows the weights distribution of the model trained with full precision (`float32`), which is a zero-centered normal distribution. The distribution after WSQ quantization shown in Figure 4b follows the original distribution but results in 8 discrete levels. We can see that it is still zero-centered, covering both negative and positive discrete levels. However, the negative levels are infeasible in physical optical devices. Thus, after post-processing is applied, the distribution (Figure 4c) is significantly changed to satisfy all physical properties. All the negative discrete levels are truncated to zero and the positive discrete levels will be quantized to the nearest physically available discrete levels. Meanwhile, we can see that GS method weight distribution shown in Figure 4d is quite different compared to WSQ after post-processing. Note that GS method trained parameters will not need any further post-processing for hardware deployment.

In conclusion, conventional quantization methods with post-processing are applicable to discrete mapping from the device level to DONN systems but suffer from not being aware of physical feasibility for different devices. However, with the proposed GS-based framework, the model is trained with full awareness of the physical devices as discrete differential mappings. Thus, the weights acquired from the GS-based framework can be deployed directly on the physical devices without post-processing and offer significant advantages in the physics-aware algorithm-hardware correlation between numerical models and physical optical systems.

4.3 Complex-valued Confidence Evaluation

We first evaluate the same DONN system structure trained with our GS-based method and the state-of-the-art DONN training methods in [28, 44]. Specifically, we aim to overcome the training limitations stated in [28] w.r.t physics limitations of shallow DONNs (≤ 3 layers). Second, we provide the confidence evaluation for DONN systems with different structure complexity, all shown in Table 1.

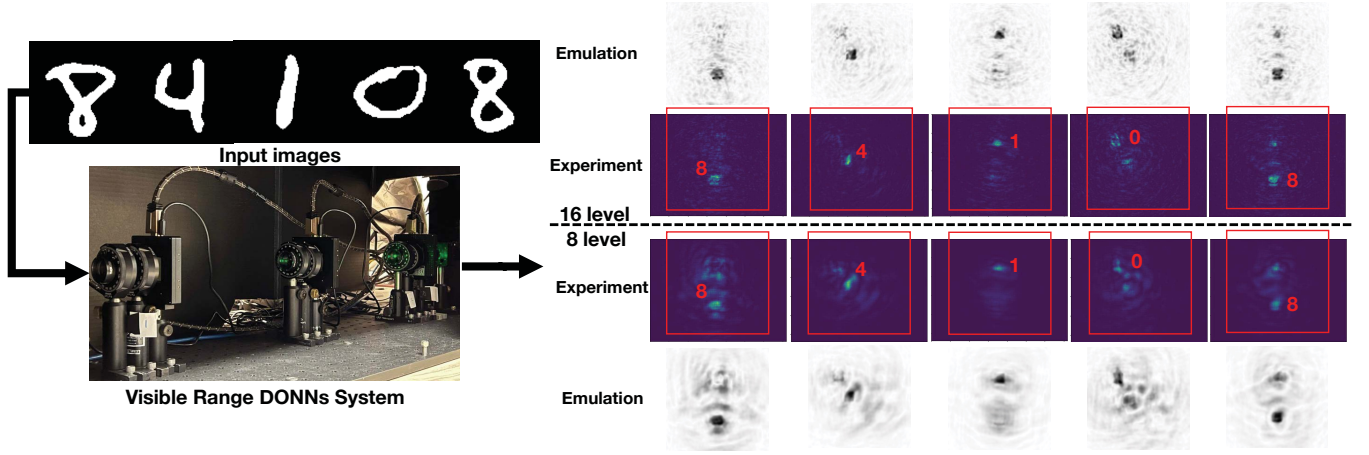


Figure 5: Experiments conducted on our existing 3-layer reconfigurable DONNs system [3] trained with our proposed approach.

First, as shown in the first column and the second column in Table 1, DONN implemented with different diffractive layers can be trained to achieve similar high prediction accuracy (0.98 for MNIST, 0.89 for FMNIST) using our framework. For the DONN systems trained in [28], the accuracy performance decreases as the DONN system structured with less diffractive layers, in which Lin et al. [28] stated that it was due to the fundamental limitations of optical physics in shallow DONNs. However, our proposed GS training framework has successfully boosted the prediction performance in those shallow architectures.

Table 1: Confidence evaluations of DONNs trained with 8 discrete values in the optical devices, with testing accuracy.

Dataset	Depths	[28, 44](0%)	0 %	1 %	3 %	5 %
MNIST	$D=1$	0.670	0.960	0.398	0	0
	$D=3$	0.910	0.978	0.961	0.876	0.661
	$D=5$	0.950	0.979	0.979	0.979	0.977
FMNIST	$D=1$	0.540	0.874	0.340	0.001	0
	$D=3$	0.830	0.889	0.791	0.518	0.278
	$D=5$	0.870	0.890	0.889	0.886	0.883

While the accuracy in shallow DONNs is significantly improved, we further analyze the performance robustness of the DONNs implemented with different depth of diffractive layers. Specifically, we explore the confidence of the predictions acquired by the system. When the sample is classified correctly, we decrease the highest probability generated by softmax function (softmax of intensity values collected in the ten detector regions) by 1%, 3% and 5%, and then evenly distribute to the other nine outputs, i.e., increasing the probabilities of the other outputs by 0.11%, 0.33% and 0.55%, respectively. We can see that for both datasets, as the depth of DONNs increases, the prediction confidence increases, i.e., the prediction accuracy drops less w.r.t the applied errors. For example, there is almost no accuracy degradation on five-layer DONNs for MNIST, and less than 1% degradation on FMNIST with up to 5% applied error. However, for single-layer DONNs, the accuracy drops 56% for MNIST and 53% for FMNIST when 1% error applied and drops to 0 when applied error increases to 3% and 5%. The confidence studies

have post great intuitions in building real-world DONN systems in different device/system precision scenarios. For example, for high precision and expensive devices and camera, shallow DONNs is sufficient; while for low precision optical system, deeper DONNs are needed for robust performance.

4.4 Physical Experimental Evaluations

The existing hardware system for verifying the demonstrated simulation results from GS-framework is shown in Figure 5 and has been presented by Chen et. al [3]. Briefly, the input images are generated at 532 nm wavelength laser and have the size 100×100 . The distance between SLMs and between the last SLM and camera is set as 27.94 cm (11 inch). The final diffraction pattern is captured on a CMOS camera. For a 3-layer DONN system for MNIST data shown in Table 1, the trained discrete weights (voltage level) of the 3-layer DONNs shown in Table 1 are directly deployed to the system that controls the applied phase modulation of each device, whose functionality follows the phase and amp function curves in Figure 1c. As shown in Figure 5, we can see that the DONNs simulation results match the experimental measurements very well, which demonstrate the effectiveness of our GS framework, especially at low finite bit precision.

5 CONCLUSION

This work studies a novel flexible device-to-system hardware-software codesign framework which enables efficient training of DONNs systems implemented with arbitrary experimental measured optical devices across the layers. Specifically, this framework realizes backpropagation through discrete parameters via Gumbel-Softmax. Our simulation results demonstrate that the DONN system optimized with the proposed framework will acquire tremendous accuracy improvements compared to the state-of-the-art quantization methods. Moreover, exploration for temperature schedule for Gumbel-Softmax in DONN system, confidence evaluation for DONN systems implemented with different layers, algorithm verification on physical optical systems are comprehensively discussed. **Acknowledgement:** This work is supported by University of Utah start-up fund and NSF-2019336 and NSF-2008144.

REFERENCES

- [1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale Machine Learning. *Symp. on Operating System Design and Implementation (OSDI)* (2016), 265–283.
- [2] Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453* (2019).
- [3] Ruiyang Chen, Yingjie Li, Minhan Lou, Jichao Fan, Yingheng Tang, Berardi Sensale-Rodriguez, Cunxi Yu, and Weilu Gao. 2022. Physics-aware Complex-valued Adversarial Machine Learning in Reconfigurable Diffractive All-optical Neural Network. *arXiv preprint arXiv:2203.06055* (2022).
- [4] Ruiyang Chen, Yingjie Li, Minhan Lou, Cunxi Yu, and Weilu Gao. 2022. Complex-valued Reconfigurable Diffractive Optical Neural Networks using Cost-effective Spatial Light Modulators. In *CLEO: Applications and Technology*. Optica Publishing Group, JTh3B–56.
- [5] Okan K Ersoy. 2006. *Diffraction, Fourier optics and imaging*. John Wiley & Sons.
- [6] J Feldmann, N Youngblood, C David Wright, H Bhaskaran, and WHP Pernice. 2019. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* 569, 7755 (2019), 208–214.
- [7] Yonggan Fu, Yonggan Zhang, Chaojian Li, Zhongzhi Yu, and Yingyan Lin. 2021. A3C-S: Automated Agent Accelerator Co-Search towards Efficient Deep Reinforcement Learning. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 13–18.
- [8] Yonggan Fu, Yonggan Zhang, Yang Zhang, David Cox, and Yingyan Lin. 2021. Auto-NBA: Efficient and effective search over the joint space of networks, bitwidths, and accelerators. In *International Conference on Machine Learning*. PMLR, 3505–3517.
- [9] Weilu Gao, Cunxi Yu, and Ruiyang Chen. 2021. Artificial Intelligence Accelerators Based on Graphene Optoelectronic Devices. *Advanced Photonics Research* 2, 6 (2021), 2100048.
- [10] Weilu Gao, Cunxi Yu, and Ruiyang Chen. 2021. Graphene Optoelectronic Artificial Intelligence Accelerators. In *CLEO: QELS_Fundamental Science*. Optical Society of America, JTu3A–88.
- [11] Jiaqi Gu, Chenghao Feng, Zheng Zhao, Zhoufeng Ying, Ray T Chen, and David Z Pan. 2021. Efficient on-chip learning for optical neural networks through power-aware sparse zeroth-order optimization. In *Proc. AAAI*.
- [12] Jiaqi Gu, Chenghao Feng, Zheng Zhao, Zhoufeng Ying, Mingjie Liu, Ray T Chen, and David Z Pan. 2021. SqueezeLight: towards scalable optical neural networks with multi-operand ring resonators. In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 238–243.
- [13] Jiaqi Gu, Zheng Zhao, Chenghao Feng, Wuxi Li, Ray T Chen, and David Z Pan. 2020. FLOPS: efficient on-chip learning for optical neural networks through stochastic zeroth-order optimization. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.
- [14] Jiaqi Gu, Zheng Zhao, Chenghao Feng, Mingjie Liu, Ray T Chen, and David Z Pan. 2020. Towards area-efficient optical neural networks: an FFT-based architecture. In *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 476–481.
- [15] Emil Julius Gumbel. 1954. *Statistical theory of extreme values and some practical applications: a series of lectures*. Vol. 33. US Government Printing Office.
- [16] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *Advances in Neural Information Processing Systems (NIPS)* (2015).
- [17] Adir Hazan, Barak Ratzker, Danzhen Zhang, Aviad Katiyi, Nachum Frage, Maxim Sokol, Yury Gogotsi, and Alina Karabchevsky. 2021. Ti_3C_2Tx MXene Enabled All-Optical Nonlinear Activation Function for On-Chip Photonic Deep Neural Networks. *arXiv preprint arXiv:2109.09177* (2021).
- [18] Chaoyang He, Haishan Ye, Li Shen, and Tong Zhang. 2020. Milenas: Efficient neural architecture search via mixed-level reformulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11993–12002.
- [19] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).
- [20] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. 2017. In-datacenter Performance Analysis of a Tensor Processing Unit. *Int'l Symp. on Computer Architecture (ISCA)* (2017), 1–12.
- [21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [22] Raghuraman Krishnamoorthi. 2018. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342* (2018).
- [23] Yann LeCun. 1998. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/> (1998).
- [24] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.
- [25] Mengquan Li, Zhongzhi Yu, Yonggan Zhang, Yonggan Fu, and Yingyan Lin. 2021. O-HAS: Optical Hardware Accelerator Search for Boosting Both Acceleration Performance and Development Speed. In *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. IEEE, 1–9.
- [26] Yingjie Li, Ruiyang Chen, Berardi Sensale Rodriguez, Weilu Gao, and Cunxi Yu. 2021. Multi-Task Learning in Diffractive Deep Neural Networks via Hardware-Software Co-design. *Scientific Reports* (2021), 1–9.
- [27] Yingjie Li and Cunxi Yu. 2021. Late Breaking Results: Physical Adversarial Attacks of Diffractive Deep Neural Networks. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1374–1375.
- [28] Xing Lin, Yair Rivenson, Nezh T Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. 2018. All-optical machine learning using diffractive deep neural networks. *Science* 361, 6406 (2018), 1004–1008.
- [29] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).
- [30] Deniz Mengü, Yi Luo, Yair Rivenson, and Aydogan Ozcan. 2019. Analysis of diffractive optical neural networks and their integration with electronic neural networks. *IEEE Journal of Selected Topics in Quantum Electronics* 26, 1 (2019), 1–14.
- [31] Deniz Mengü, Yair Rivenson, and Aydogan Ozcan. 2020. Scale-, shift- and rotation-invariant diffractive optical networks. *arXiv preprint arXiv:2010.12747* (2020).
- [32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.
- [33] Md Sadman Sakib Rahman, Jingxi Li, Deniz Mengü, Yair Rivenson, and Aydogan Ozcan. 2020. Ensemble learning of diffractive optical networks. *arXiv preprint arXiv:2009.06869* (2020).
- [34] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Zidek, Alexander WR Nelson, Alex Bridgland, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577, 7792 (2020), 706–710.
- [35] Hardik Sharma, Jongse Park, Emmanuel Amaro, Bradley Thwaites, Praneetha Kotha, Anmol Gupta, Joon Kyung Kim, Asit Mishra, and Hadi Esmaeilzadeh. 2016. Dnnweaver: From high-level deep network models to fpga acceleration. In *the Workshop on Cognitive Architectures*.
- [36] Yichen Shen, Nicholas C Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, et al. 2017. Deep learning with coherent nanophotonic circuits. *Nature Photonics* 11, 7 (2017), 441.
- [37] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.
- [38] Alexander N Tait, Thomas Ferreira De Lima, Ellen Zhou, Allie X Wu, Mitchell A Nahmias, Bhavin J Shastri, and Paul R Prucnal. 2017. Neuromorphic photonic networks using silicon photonic weight banks. *Scientific reports* 7, 1 (2017), 1–10.
- [39] Karen Ullrich, Edward Meeds, and Max Welling. 2017. Soft weight-sharing for neural network compression. *arXiv preprint arXiv:1702.04008* (2017).
- [40] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. 2019. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10734–10742.
- [41] Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, and Kurt Keutzer. 2018. Mixed precision quantization of convnets via differentiable neural architecture search. *arXiv preprint arXiv:1812.00090* (2018).
- [42] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).
- [43] Zhoufeng Ying, Chenghao Feng, Zheng Zhao, Shounak Dhar, Hamed Dalir, Jiaqi Gu, Yue Cheng, Richard Soref, David Z Pan, and Ray T Chen. 2020. Electronic-photonic arithmetic logic unit for high-speed computing. *Nature communications* 11, 1 (2020), 1–9.
- [44] Tiankuang Zhou, Xing Lin, Jiamin Wu, Yitong Chen, Hao Xie, Yipeng Li, Jingtao Fan, Huaqiang Wu, Lu Fang, and Qionghai Dai. 2021. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nature Photonics* 15, 5 (2021), 367–373.