ROQ: A Noise-Aware Quantization Scheme Towards Robust Optical Neural Networks with Low-bit Controls

Jiaqi Gu[1], Zheng Zhao[1], Chenghao Feng[1], Hanqing Zhu[2], Ray T. Chen[1], and David Z. Pan[1]

[1]ECE Department, The University of Texas at Austin
[2]Department of Microelectronics, Shanghai Jiao Tong University, China
{jqgu, zhengzhao, fengchenghao1996}@utexas.edu, chenrt@austin.utexas.edu, dpan@ece.utexas.edu

***Abstract*— Optical neural networks (ONNs) demonstrate orders-of-magnitude higher speed in deep learning acceleration than their electronic counterparts. However, limited control precision and device variations induce accuracy degradation in practical ONN implementations. To tackle this issue, we propose a quantization scheme that adapts a full-precision ONN to low-resolution voltage controls. Moreover, we propose a protective regularization technique that dynamically penalizes quantized weights based on their estimated noise-robustness, leading to an improvement in noise robustness. Experimental results show that the proposed scheme effectively adapts ONNs to limited-precision controls and device variations. The resultant four-layer ONN demonstrates higher inference accuracy with lower variances than baseline methods under various control precisions and device noises.**

I. Introduction

As Moore's Law slows down, the optical neural network (ONN) offers a promising alternative to its electronic counterpart due to ultra-low latency and high energy efficiency [1]–[3]. An integrated fully-optical neural network [1] consisting of Mach-Zehnder Interferometer (MZI) arrays has been demonstrated to perform matrix multiplication based on matrix singular value decomposition (SVD) and unitary parametrization [4], [5], with over 100 GHz photo-detection rate and near-zero energy dissipation [6].

However, similar to other neuromorphic systems [7], [8], ONNs inevitably bear a challenge in robustness to non-ideal effects in the actual implementation. First, phase shifts produced by MZIs are not physically implementable with arbitrary precision since the electronic control of optical devices only has limited resolution. Therefore, weight encoding errors exist when mapping full-precision models onto physical optical devices. Also, low-precision voltage controls are preferred in neuromorphic platforms [9], [10] for energy, performance, and control complexity considerations. Thus we are motivated to put forward an effective methodology to design ONNs adaptive to low-bit controls. Another critical issue is the device-level noise on MZIs. Each MZI contains a configurable thermo-optic phase shifter to encode the ONN weight. This phase shift can be influenced by the device size, manufacturing imperfection, voltage control, and environmental changes, etc [1], [11], inducing weight encoding errors. Given the cascaded architecture of ONNs, phase errors caused by limited control resolution and phase shifter variations will propagate and accumulate throughout the entire system, eventually degrade the inference accuracy [1].

Recently, most ONN research focuses on new devices and novel architectures, targeting at the area and power improvement [2], [12], while limited works investigate robustness issues of ONNs. A proposed FFT-style architecture is demonstrated to have better robustness to device imperfections than the original ONN [13]. Another work [2] proposed a slimmed ONN architecture which cuts down the number of MZIs using a new decomposition method to eliminate part of noise sources. The above two previous works demonstrate novel ONN architectures with better robustness, but require special hardware implementations and ideally assume full precision controls. We are the first to propose the design methodology that addresses low control precision and device variations of ONNs.

Therefore, we propose a noise-aware quantization scheme *ROQ* to help design a robust ONN model that is amenable to low-precision controls and device variations. The main contributions of this work are as follows,

- We experimentally show that naive post-training quantization and traditional iterative quantization methods perform poorly on ONN voltage-domain discretization and can barely improve its noise robustness.
- We propose an end-to-end quantization scheme to enable low-precision voltage control of ONNs and mitigate the corresponding accuracy degradation.
- A protective Group Lasso regularization technique is proposed to boost noise-robustness of quantized ONNs.

II. Preliminaries

In this section, we introduce the background knowledge for our proposed ONN training methodology.

A. ONN Architecture

The classical integrated ONN architecture [1] implements Mach-Zehnder Interferometer (MZI) arrays to realize MLP inference, shown in Fig. 1. This ONN architecture first decomposes the weight matrix $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ through singular value decomposition (SVD) $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ and then performs unitary parametrization to $\boldsymbol{U}$ and $\boldsymbol{V}^*$ as follows [4],

$$\boldsymbol{U}(n) = \boldsymbol{D}\prod_{i=n}^{2}\prod_{j=1}^{i-1}\boldsymbol{R}_{ij}, \tag{1}$$

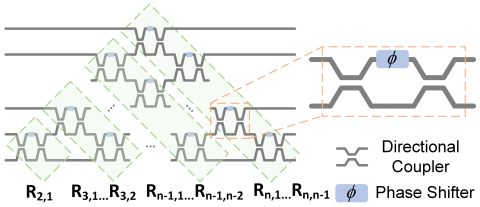


Fig. 1: Schematic of a triangular MZI array and the structure of a 2×2 MZI.

where $\boldsymbol{D}$ is an n-dimensional diagonal matrix that only contains ± 1, and the planar rotator $\boldsymbol{R}_{ij}$ is an $n \times n$ identity matrix, where four entries at (i,i), (i,j), (j,i) and (j,j) indices are replaced by $\cos\phi$, $\sin\phi$, $-\sin\phi$, and $\cos\phi$. Each $\boldsymbol{R}_{ij}$ can be implemented with a 2×2 MZI, whose transfer function is:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{2}$$

where the phase ϕ can be implemented with a optical phase shifter. The diagonal matrix $\boldsymbol{\Sigma}$ can be realized by optical attenuators or amplification materials to perform signal scaling.

B. Neural Network Quantization

Extensive works have shown that more efficient DNNs can be achieved by low-bit parameter quantization. However, post-training quantization error can lead to performance degradation. In order to compensate the accuracy loss, a straightforward iterative method that performs quantization and re-training alternately can be used to reduce the quantization error. However, it may encounter divergence issues when weights are quantized with very low precision. Another successful approach is to perform quantization-aware training with back-propagation [14]. To achieve that, it requires to define a mechanism for gradient propagation through the non-differentiable quantization operation. Typically, a straight-through estimator (STE) [15] is used to model the gradient of discretization, such that the output gradient can directly back-propagate to its input as,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_q} \odot \mathbb{1}_{\boldsymbol{W} \in [minV, maxV]}, \tag{3}$$

where $\odot$ is element-wise multiplication to cancel the gradient where $\boldsymbol{W}$ exceeds the pre-defined valid range.

III. Noise-Aware Quantization Scheme

We first analyze two major non-ideal effects of ONNs, i.e., limited resolution in voltage control and gamma noise in phase shifters in Section III-A. Then we demonstrate our proposed noise-aware quantization scheme in details. Figure 2 shows the framework of the proposed *ROQ* scheme.

A. Limited Control Resolution and Phase Shifter Gamma Noise

In the ONN architecture discussed in Section II-A, the phase shifter is controlled by an electronic signal. The relation between voltage control v and the configured phase shift ϕ is typically modeled as $\phi = \gamma v^2$ [1], [16], where γ is a device-level coefficient that can be calculated by $\gamma = \pi / v_\pi^2$, and v_π is defined as the voltage required to achieve π radians phase shift. Limited by the precision of voltage supply, e.g., b-bit with dynamic range $[0, v_{max}]$, only $\lfloor \frac{v_{2\pi}}{v_{max}} 2^b \rfloor$ non-uniformly discretized phase values are achievable. We show the 6-bit quantized distribution of an arbitrary 64×64 unitary matrix in Fig. 3. The interval between two phase levels is quadratically enlarged as the voltage increases, leading to a larger phase encoding error.

However, as the γ coefficient is affected by manufacturing variations and the environmental changes, the actual phase shifts will be perturbed by the gamma noise Given the quadratic relation between ϕ and v, larger phase shifts are more sensitive to gamma noise. In this paper, we model this gamma noise as a random variable sampled from a Gaussian distribution $\Delta\gamma \sim \mathcal{N}(0, \sigma^2)$.

B. Voltage-Domain Quantization

This section will focus on the detailed quantization method applied in the *voltage domain*, which is the major difference between this paper and previous weight quantization works. In this paper, we adopt blocking matrix multiplications to implement fully-connected layers for tractable weight encoding error and practical considerations [1], [17], [18]. The weight matrix $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ is partitioned into $p \times q$ sub-matrices, where each $k \times k$ block is realized by an MZI array.

A naive method is to directly train quantized phases with back-propagation algorithm, However, gradient propagation through Eq.(1) is extreme inefficient. Besides, analytically computing the gradient of unitary matrices w.r.t each phase also casts daunting difficulties on computations [1]. To resolve the challenge, we propose a projected quantization method with coarse gradient approximation to perform voltage-domain discretization. The entire procedure starts from training a full-precision ONN. Then, *ROQ* scheme is performed to fine-tune the decomposed matrices $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. We first partition the weight matrix into a batch of square blocks $\boldsymbol{W} \in \mathbb{R}^{p \times q \times k \times k}$, each $k \times k$ block denoted as $\boldsymbol{W}_{ij}$, then we use the decompositions $\boldsymbol{U}_{ij}\boldsymbol{\Sigma}_{ij}\boldsymbol{V}_{ij}^*$ of each $\boldsymbol{W}_{ij}$ as initialization. Each optimization step consists of two stages as shown in Fig. 2. The first stage efficiently propagates coarse gradient through parametrization and quantization. A subsequent unitary projection stage maps the updated matrices back to unitary sub-spaces to meet the orthogonality constraint on $\boldsymbol{U}$ and $\boldsymbol{V}^*$. For brevity, we only focus on one unitary block $\boldsymbol{U}_{ij}$, simplified as $\boldsymbol{U}$.

1) Coarse Gradient Approximation: In this stage, we model the voltage-domain quantization as a straight-through estimator (STE) [15] to approximate its undefined derivative. Specifically, in the forward propagation, $\boldsymbol{U}^t$ is parametrized into $\boldsymbol{\Phi}^t$ and $\boldsymbol{v}^t$ based on Eq. (1) and $\phi = \gamma v^2$. Then, we perform quantization to get discretized voltages $\boldsymbol{v}_q^t$ with the resolution $v_{max}/(2^b - 1)$. Any quantized voltage that exceeds the valid range $[0, v_{2\pi})$ is processed with a clipping function with phase wrapping to guarantee the validity constraint,

$$v_{q,c} = \texttt{WrapClip}(v_q) = \begin{cases} v_q, & \text{if } 0 \le v_q < v_{2\pi} \\ 0, & \text{if } v_q \ge v_{2\pi}. \end{cases} \tag{4}$$

Invalid large voltages are clipped to 0 instead of their closest valid quantization level because smaller phases have less quantization errors and more quantization levels are distributed around small phase values such that better model expressivity can be maintained. Clipped voltages $\boldsymbol{v}_{q,c}^t$ are used to reconstruct $\boldsymbol{\Phi}_q^t$

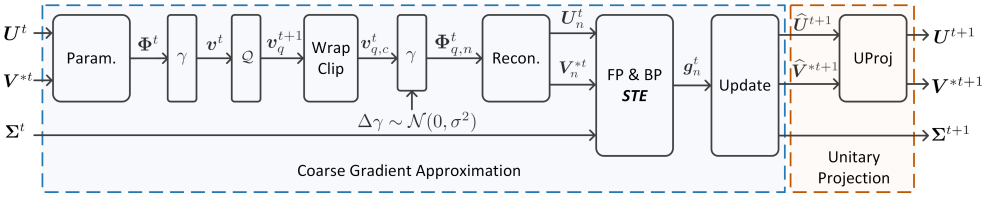
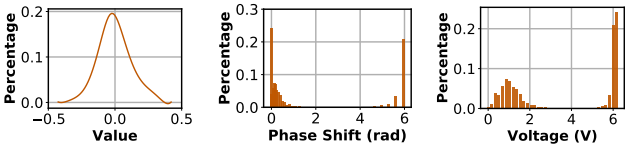


Fig. 2: Framework of ROQ flow at one optimization step. *FP & BP* represents forward and backward propagation; *Param.*, *Recon.*, and *STE* are short for parametrization, reconstruction, and straight-through estimator respectively.



(a) (b) (c)

Fig. 3: Histograms of (a) 6-bit quantized unitary matrix, (b) 6-bit quantized phases, (c) 6-bit quantized voltages. The dynamic range of the voltage supply is assumed to be [0, 10.8V], thus the resolution is 171.4 mV.

and $\boldsymbol{U}_q^t$ using Eq. (1). All subsequent forward computations are based on quantized unitary matrices. In the backward propagation, we coarsen the whole b-bit quantization process $\boldsymbol{U}_q^t = \mathcal{Q}_b(\boldsymbol{U}^t)$ as an entirety and efficiently estimate its coarse gradient with an STE. The gradient propagation of $\mathcal{Q}$ follows $g_q^t = \frac{\partial L^t}{\partial U^t} = \frac{\partial L^t}{\partial U_q^t}$. In this way, the gradient can efficiently propagate through $\mathcal{Q}$ without computing the complicated gradient inside of it, and we denote the updated unitary matrix as $\widehat{\boldsymbol{U}}^{t+1}$.

2) Unitary Projection: The orthogonality of matrix $\boldsymbol{U}^t$ is a prerequisite of unitary parametrization, but the above gradient descent method inevitably drives $\widehat{\boldsymbol{U}}^{t+1}$ to an infeasible point. This hard constraint of orthogonality can be satisfied through unitary projection $\boldsymbol{U} = \texttt{UProj}(\widehat{\boldsymbol{U}})$ defined as,

$$\boldsymbol{P}\boldsymbol{S}\boldsymbol{Q}^* = \texttt{SVD}(\widehat{\boldsymbol{U}}); \quad \boldsymbol{U} = \boldsymbol{P}\boldsymbol{Q}^*. \tag{5}$$

We project $\widehat{\boldsymbol{U}}^{t+1}$ back into unitary planes to get $\boldsymbol{U}^{t+1}$ with minimum approximation error [2] at each iteration.

Our proposed method has the sequential time complexity of $\mathcal{O}(pqk^3) = \mathcal{O}(kn^2)$, which is computationally efficient to train quantized ONNs. This quadratic complexity attributes to the coarse gradient approximation algorithm which does not require the gradient propagation through parametrization. This enables an efficient implementation of Eq. (1) without expensive matrix multiplication. Each time an $\boldsymbol{R}_{ij}$ left-multiplies a matrix, it is equivalent to update two rows of the matrix,

$$\boldsymbol{R}_{ij} \cdot \boldsymbol{W} \Longleftrightarrow \begin{cases} \boldsymbol{W}(i,:) = \cos\phi_{ij}\ \boldsymbol{W}(i,:) + \sin\phi_{ij}\ \boldsymbol{W}(j,:) \\ \boldsymbol{W}(j,:) = \cos\phi_{ij}\ \boldsymbol{W}(j,:) - \sin\phi_{ij}\ \boldsymbol{W}(i,:). \end{cases} \tag{6}$$

Note that our method can perform parametrization to a batch of unitary blocks in parallel with even lower runtime cost.

C. Noise-Aware Training with Protective Group Lasso Regularization

To further augment the robustness of ONNs, we propose a protective regularization strategy based on dynamic evaluation on the noise robustness. For less robust weight matrix blocks, i.e., with a larger error $||\boldsymbol{W} - \boldsymbol{W}_n||_2^2$, we will diminish its significance by exerting larger penalty on its ℓ_2 norm.

Based on this robustness estimation, we propose a protective Group Lasso regularization loss function to guide the ONN model to more robust local optima as follows,

$$\mathcal{L}_{PGL} = \sum_{l,i,j=1}^{L,p^l,q^l} \underbrace{\frac{||\boldsymbol{W}_{ij,q}^l - \boldsymbol{W}_{ij,q,n}^l||_2^2}{\max_{i,j} ||\boldsymbol{W}_{ij,q}^l - \boldsymbol{W}_{ij,q,n}^l||_2^2}}_{P_{ij}^l} \sqrt{1/\beta_{ij}^l}\ ||\boldsymbol{W}_{ij}^l\ ||_2^2, \tag{7}$$

where L is the number of layers in the ONN model; p^l, q^l are number of blocks along output and input channels in the l-th layer; the last term is group-wise ℓ_2 norm; and β_{ij}^l is the number of elements in $\boldsymbol{W}_{ij}^l$ to balance different group sizes. P_{ij}^l is the protective coefficient used to estimates the robustness of a quantized block $\boldsymbol{W}_{ij,q}^l$, which is obtained by gamma noise injection during forward propagation. This coefficient normalizes the error in each ONN layer to $[0, 1]$, such that less robust block will suffer larger Group Lasso penalty and thus the resultant weight matrix is protected from gamma noise perturbation. To obtain a more stable estimation of noise robustness we apply exponential moving average (EMA) to make this coefficient learnable,

$$\widehat{P}_{ij}^{l(t)} = \eta \widehat{P}_{ij}^{l(t-1)} + (1-\eta) P_{ij}^{l(t)}, \tag{8}$$

where the adaptivity rate η is set to 0.999. This EMA-based method efficiently estimates a more stable and accurate protective coefficient $\mathbb{E}_{\Delta\gamma\sim\mathcal{N}(0,\sigma^2)}[P_{ij}^l]$ as it converges.

However, evaluating the protective coefficients with a constant intensity of noise comes at the cost of a loss in accuracy when the noise is large. Thus we adopt a noise source cooling strategy to periodically decay the noise standard deviation σ to help convergence.

IV. Experimental Results

To evaluate the effectiveness of our proposed noise-aware quantization scheme, we compare the inference accuracy and noise robustness with post-training quantization and iterative quantization methods on the MNIST dataset [19]. Without loss of generality, we use a study case of a four-layer ONN with configuration of (12×12)-64(8)-64(8)-40(10)-10, where 64(8) denotes 64 output channels with size-8 blocks. For voltage quantization, we set up the same parameters used in the original ONN architecture [1], where $v_{max} = 10.8V$ and $v_\pi = 4.36V$.

A. Comparison under Limited Voltage Resolution

Starting from a full-precision ONN with 97.6% accuracy, we compare the effectiveness of 1) naive post-training quantization (`Naive`), 2) iterative quantization (`Baseline`), and 3) our proposed method (`ROQ`). The `Naive` method directly quantizes the voltage controls of the full-precision ONN. The `Baseline`

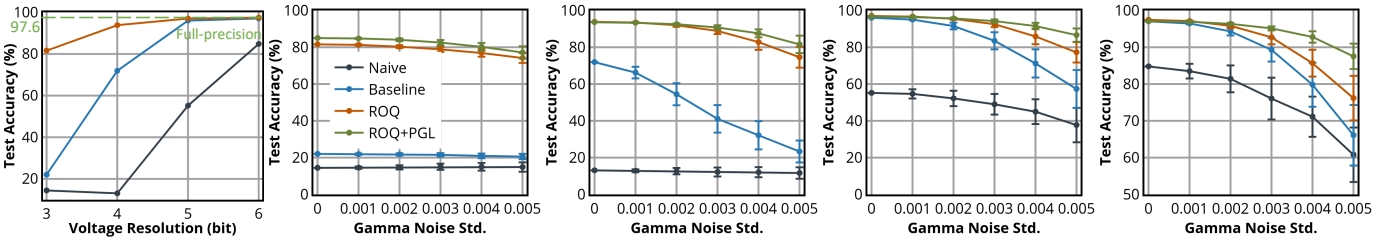


Fig. 4: (a) Comparison among different quantization methods. (b), (c), (d), (e) Robustness comparison with 3-, 4-, 5-, 6-bit voltage controls respectively. Error bars on the line represent the $\pm 1 \cdot \sigma$ accuracy variance for 100 noise samples.

method performs alternating steps of voltage-domain quantization and re-training to recover accuracy loss until convergence.

Figure 4a illustrates that our proposed method outperforms other two methods under 3- to 6-bit voltage control resolutions with much less accuracy degradation. While `Naive` and `Baseline` methods suffer from severe accuracy loss, our proposed `ROQ` method achieves $> 80\%$ inference accuracy with 3-bit voltage controls and $\sim 97\%$ accuracy on higher bits.

B. Comparison under Phase Shifter Gamma Noise

Under low-precision voltage control, we further inject gamma noise in the MZI arrays during inference with various intensities and evaluate the noise robustness of 1) post-training quantization method (`Naive`), 2) iterative quantization method (`Baseline`), 3) the proposed method (`ROQ`), and 4) ROQ with protective Group Lasso regularization (`ROQ+PGL`). On four different bit widths, from 3 to 6 bit, we evaluate their inference accuracy with five different noise variances, from σ=0.001 to 0.005. All statistics are averaged by randomly sampling 100 noise samples. Figure 4b- 4e show the mean inference accuracy and $\pm 1 \cdot \sigma$ uncertainty of all comparison methods.

Our proposed `ROQ` outperforms the `Naive` and `Baseline` methods on various noise intensities. When the noise standard deviation reaches 0.005, it still achieves $\sim 80\%$ test accuracy. Besides, `ROQ` leads to a smaller accuracy variance (narrower error bar) than the `Baseline` method over 100 noise samples. Assisted by the `PGL` technique, `ROQ` demonstrates even better gamma noise tolerance. The proposed `ROQ` scheme and `PGL` technique enable an end-to-end flow to fine-tune a full-precision ONN model and make it adaptive to low-precision controls and phase shifter gamma noise. The resultant ONN raises the accuracy from $\sim 20\%$ (`Baseline`) to $\sim 80\%$ with lower variances on a downsampled MNIST dataset under 3-bit voltage control resolution together with a relatively large gamma noise $\sigma = 0.005$ in phase shifters.

V. Conclusion

In this work, we propose a training methodology to adapt ONNs to low-precision controls and non-ideal environment with phase shifter noises. Our proposed ROQ performs an end-to-end ONN fine-tuning with discretized voltage controls via coarse gradient approximation and unitary projection. A protective Group Lasso (PGL) regularization technique is also proposed to protect ONNs from phase shifter noises by dynamically suppressing less robust weight matrix blocks. Experimental results show that, compared with the baseline method, the proposed PGL-assisted ROQ can effectively tackle the non-ideal issues of ONNs and provide a low-overhead approach towards noise-robust ONN accelerators with lower control complexity.

Acknowledgment

The authors acknowledge the Multidisciplinary University Research Initiative (MURI) program through the Air Force Office of Scientific Research (AFOSR), contract No. FA 9550-17-1-0071, monitored by Dr. Gernot S. Pomrenke.

References

[1] Y. Shen, N. C. Harris, S. Skirlo *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, 2017.

[2] Z. Zhao, D. Liu, M. Li *et al.*, "Hardware-software co-design of slimmed optical neural networks," in *Proc. ASPDAC*, 2019.

[3] J. Gu, Z. Zhao, C. Feng *et al.*, "Towards area-efficient optical neural networks: an FFT-based architecture," in *Proc. ASPDAC*, 2020.

[4] M. Reck, A. Zeilinger, H. Bernstein *et al.*, "Experimental realization of any discrete unitary operator," *Physical review letters*, 1994.

[5] A. Ribeiro, A. Ruocco, L. Vanacker *et al.*, "Demonstration of a 4×4-port universal linear circuit," *Optica*, 2016.

[6] L. Vivien, A. Polzer, D. Marris-Morini *et al.*, "Zero-bias 40gbit/s germanium waveguide photodetector on silicon," *Opt. Express*, 2012.

[7] Z. He, J. Lin, R. Ewetz *et al.*, "Noise injection adaption: End-to-end reram crossbar non-ideal effect adaption for neural network mapping," in *Proc. DAC*, 2019.

[8] A. S. Rekhi, B. Zimmer, N. Nedovic *et al.*, "Analog/mixed-signal hardware error modeling for deep learning inference," in *Proc. DAC*, 2019.

[9] L. Song, X. Qian, H. Li *et al.*, "Pipelayer: A pipelined reram-based accelerator for deep learning," 2017.

[10] M. Hu, J. P. Strachan, Z. Li *et al.*, "Dot-product engine for neuromorphic computing: Programming 1t1m crossbar to accelerate matrix-vector multiplication," in *Proc. DAC*, 2016.

[11] N. C. Harris, Y. Ma, J. Mower, T. Baehr-Jones, D. Englund, M. Hochberg, and C. Galland, "Efficient, compact and low loss thermo-optic phase shifter in silicon," *Opt. Express*, 2014.

[12] R. Hamerly, L. Bernstein, A. Sludds *et al.*, "Large-scale optical neural networks based on photoelectric multiplication," *Phys. Rev. X*, 2019.

[13] M. Y. S. Fang, S. Manipatruni, C. Wierzynski *et al.*, "Design of optical neural networks with component imprecisions," *Optics Express*, 2019.

[14] I. Hubara, M. Courbariaux, D. Soudry *et al.*, "Quantized neural networks: Training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, 2017.

[15] G. Hinton, "Neural networks for machine learning," *Coursera Video Lecture*, 2012.

[16] E. Timurdogan, Z. Su, C. V. Poulton *et al.*, "Aim process design kit (aimpdkv2.0): Silicon photonics passive and active component libraries on a 300mm wafer," in *Optical Fiber Communication Conference*, 2018.

[17] Y. Ji, Y. Zhang, S. Li *et al.*, "NEUTRAMS: Neural network transformation and co-design under neuromorphic hardware constraints," in *Proc. MICRO*, 2016.

[18] Y. Ji, Y. Zhang, W. Chen *et al.*, "Bridge the gap between neural networks and neuromorphic hardware with a neural network compiler," in *Proc. ASPLOS*, 2018.

[19] Y. LeCun, "The MNIST database of handwritten digits," http://yann.lecun.com/exdb/mnist/, 1998.