Adversarial Attacks on an Optical Neural Network

Shuming Jiao, Ziwei Song, and Shuiying Xiang

Abstract—Adversarial attacks have been extensively investigated for machine learning systems including deep learning in the electronic domain. However, the adversarial attacks on optical neural networks (ONNs) have been seldom considered previously. In this work, we first construct an accurate image classifier with an ONN using a mesh of interconnected Mach–Zehnder interferometers (MZIs). Then a corresponding adversarial attack scheme is proposed for the first time. The attacked images are visually very similar to the original ones but the ONN system becomes malfunctioned and generates wrong classification results in most time. The results indicate that adversarial attack is also a significant issue for optical machine learning systems.

Index Terms—Adversarial attack, optical neural network, optical computing, Mach-Zehnder interferometer, machine learning.

I. INTRODUCTION

In the modern information era, the explosive growth of data every day is a critical issue and efficient computer technology for enormous data processing is always an urgent demand. The dominating electronic computing technology is currently facing various challenges and bottlenecks despite the great success it has achieved in the past few decades. In recent years, optical computing has received much research attention [1], [2] as an alternative approach. Compared with electronic computing, optical computing has potential advantages of high-speed parallel processing and low power consumption.

The rapid development of artificial intelligence (AI) has a significant impact on many applications such as autonomous driving, smart healthcare, robotics and natural language processing. Deep learning neural network is a representative AI model by simulating the working mechanism of massive interconnected neurons in a human brain. A deep learning neural network can be conventionally implemented by electronic hardware including Central Processing Unit (CPU), Graphics Processing Unit (GPU) and Field Programmable Gate Array (FPGA). The realization of optical neural network (ONN) has recently become a crucial concern for the optical computing community as well.

Various ONN schemes in analogy to electronic deep learning have been proposed such as cascaded Mach–Zehnder interferometer (MZI) network [3], [4], [5], [6], [7], optical diffractive neural network [8], [9], [10], [11], [12] and wavelength multiplexing network [13]. Among them, an ONN with cascaded MZIs [3], [4], [5], [6], [7] has advantages of compact system size and good compatibility with silicon photonics. On the one hand, ONNs demonstrate the capability of performing similar intelligent tasks as electronic neural networks, with potentially better efficiency. For example, an ONN with cascaded MZIs can be used for machine learning tasks such as voice and image classification [3], [4], [5], [6], [7]. Two types of neural networks (i.e., optical ones and electronic ones) have many common properties such as weighted summation of signals from previous layers of neurons, better performance with more layers (depth advantage) [12], backpropagation training algorithms [4] and dropout operation for simplifying the network structure [10]. On the other hand, ONNs exhibit some unique features compared with electronic neural networks. For example, the all-optical implementation of a nonlinear activation function in a neural network is more difficult than that in the electronic domain [14].

For electronic machine learning algorithms including deep learning, the concept of adversarial attack has been extensively investigated [15], [16], [17], [18], [19], [20], [21], [22], [23]. As an interesting phenomenon, adversarial attacks refer to perturbing input images subtly with minor changes for intentionally generating wrong classification results by “fooling” the machine learning system. They pose general and serious threats to machine learning and the revealed system vulnerability has profound implications. Adversarial attacks can be divided into white-box attacks and black-box attacks [16]. In the former scenario, it is assumed that the parameters, principles, and structures of the machine learning system are known. Attackers can carry out adversarial attacks based on all the known information. In the latter scenario, the attack is realized indirectly. Adversarial samples are generated usually based on direction sensitivity estimation. The objective of direction sensitivity estimation is to obtain the dimensions of input samples for a machine learning system that are most sensitive in affecting output results with minimal variation. Then conventional samples can be slightly modified in these dimensions for becoming adversarial samples. Common estimation algorithms [16] include convex optimization, gradient-based methods, probability-based methods, differential evolution etc.
Due to the similarity between optical and electronic neural networks, adversarial attacks are likely to be general phenomena for many ONNs as well. However, the adversarial attack on an ONN was seldom noticed and considered in the past. In some previous works [24], [25], [26], optical imaging systems are used to generate or prevent adversarial attack for an electronic deep learning system. These works are closely related to adversarial attacks but the attacking target is not an optical machine learning system. In this paper, adversarial attacks are performed on a major type of ONN, i.e., a MZI mesh system [3], [4], [5], [6], [7], for the first time.

This paper is structured as follows. In Section II, the basic principles of constructing an image classifier with an ONN consisting of cascaded MZIs are described. Our proposed adversarial attack scheme regarding to this ONN system is introduced. In Section III, the proposed adversarial attack scheme is verified by comprehensive simulation results. The novelty and significance of our work is further analyzed and discussed. In Section IV, a brief conclusion is made.

II. PRINCIPLES OF PROPOSED SCHEME

In an ONN with cascaded MZIs, each MZI unit has two input ports and two output ports, shown in Fig. 1(a). A MZI mesh consists of many interconnected MZIs. The entire MZI mesh can model a vector-matrix multiplication operation. For example, an input vector consisting of 4 elements is multiplied with a 4 × 4 matrix and the output vector consists of 4 elements. Then the cascaded MZI system will have 4 input ports and 4 output ports respectively. The intensities of coherent light signals at different input or output ports correspond to the vector element values. Each individual MZI unit has two phase shifters and their phase values can be flexibly adjusted to control the linear matrix transformation between the complex-amplitudes of input and output signals, shown in Fig. 1(a). The signals at two input ports are denoted by \( e_1 \) and \( e_2 \). The signals at two output ports are denoted by \( f_1 \) and \( f_2 \). Each beam splitter will divide the input signal from the upper port (e.g., \( e_1 \) for the first beam splitter) into two halves with a \( \pi/2 \) phase shift. The two split signals will enter the upper and lower output ports respectively. The same applies to the input signal from the lower port for each beam splitter. Controlled by the two beam splitters, the two output signals \( f_1 \) and \( f_2 \) are both constructive or destructive interference results of input signals \( e_1 \) and \( e_2 \). The two phase shifters will further contribute phase modulation freedom in the mathematical model. So the final input-output relationship of each individual MZI unit is represented by the formula in Fig. 1(a). A mesh with a sufficient number of MZI units can model an arbitrary weighting matrix precisely if all the phase shifters are assigned with optimized values.

A linear classifier for sorting images consisting of M pixels into N categories can be implemented based on vector-matrix multiplication, shown in Fig. 2. If the input vector \( X \) represents pixel values of one input image, the classification result can be indicated by the maximum value in the output vector \( Y \). Based on many training samples (different images with known classification results), the element values of weighting matrix \( W \) can be optimized. Each matrix element can be updated iteratively by the following formulas (1) and (2) for each \( i (1 \leq i \leq N) \) and \( j (1 \leq j \leq M) \) with a learning rate \( \mu \). The optimization is based on gradient decent, Softmax function and cross-entropy loss [27].

\[
T_i = \left\{ \begin{array}{ll}
\frac{\exp(y_i)}{\sum_{i=1}^{N} \exp(y_i)} & X: \text{not } i-th \text{ category} \\
\frac{\exp(y_i)}{\sum_{i=1}^{N} \exp(y_i)} - 1, & X: i-th \text{ category}
\end{array} \right.
\]

After training, all the elements in \( W \) can be further normalized to \([01]\) to ensure that all the output vector element values are...
Fig. 3. Matrix decomposition into 4 × 4 submatrices for vector-matrix multiplication with a basic block of MZI networks by temporal multiplexing.

The image classifier can be implemented by an ONN with cascaded MZIs in the following way. First, a basic block of MZI network can be constructed for a 4 × 4 submatrix. An arbitrary matrix can be decomposed into three matrices by singular value decomposition (SVD): unitary matrix U, diagonal matrix Σ, and the other unitary matrix V. As shown in Fig. 1(b), a MZI network is composed of three parts to implement three matrices respectively. There is a total of 16 MZI units in the MZI mesh in which each MZI can be reconfigured individually. The phase values of each MZI in a 4 × 4 MZI network can be optimized based on three successive multiplications [28], [29].

The vector-matrix multiplication for the image classifier above can be implemented by a temporal multiplexing of many basic blocks of 4 × 4 MZI networks, and the decomposition of calculation tasks is shown in Fig. 3. Each row of the matrix \( W_{N×M} \) is divided into groups of every 16 elements and each group is rearranged as a 4 × 4 matrix. Thus, the matrix \( W_{N×M} \) is decomposed into a total of \((M/16) \times N\) submatrices and each row corresponds to \( M/16 = 49 \) submatrices. That means the 4 × 4 MZI network will be used \((M/16) \times N\) times by reconfiguring different phase values. As shown in Fig. 3, each submatrix receives a vector of four input elements at each time. The vector-matrix multiplication results of the 4 × 4 submatrices with four different sets of 4 × 1 input vectors are accumulated to obtain one of the elements in the output vector \( y_N \) corresponding to the entire N-th row in the matrix, i.e., \( y_N = y_{N(1)} + y_{N(2)} + y_{N(3)} + y_{N(4)} \).

For other elements in the output vector, the calculation processes are the same as described above. Finally, the output vector \( Y = [y_1, y_2, \ldots, y_N]^T \) can be obtained by a cascaded MZI system.

Our proposed adversarial attack scheme against the ONN image classifier is described below. It shall be noticed that the ONN system remains fixed and unchanged in the whole process. Only the input test image is modified at a minimal level. As stated above, the maximum element in the output vector \( Y \) indicates which one of the \( N \) categories the input vector (or the input image) belongs to. It is assumed that \( X \) is originally correctly identified as the k-th category by this system \((y_k \text{ is maximum})\). The objective is to find an attacked image visually similar to the original image but it will be mis-identified as another category different from k-th one. In the original output vector, it is assumed that the second largest element is \( y_l \). \( \tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_M]^T \) denotes the attacked image and the objective function \( f(\tilde{x}_m) \) \((1 \leq m \leq M)\) for minimization is given by Eq. (3) where \( \alpha, \beta, \gamma \) are weighting coefficients. The first term is to suppress \( y_k \) corresponding to the original correct category, the second term is to boost \( y_l \) corresponding to a wrong category and the third term is to ensure that the attacked image is still similar to the original one.

\[
f(\tilde{x}_m) = \sum_{i=1}^{M} \left( \alpha w_{km} \tilde{x}_m - \beta w_{lm} \tilde{x}_m + \gamma |\tilde{x}_m - x_m|^2 \right)
\]

An optimized solution of \( \tilde{X} \) can be obtained by gradient descent and minimization of mean square error with formulas (4) and (5) where \( \mu' \) is the learning rate. The elements in \( \tilde{X} \) are initially set to be random before iterative optimization.

\[
\frac{df}{dx_m} = \alpha w_{km} - \beta w_{lm} + 2\gamma (\tilde{x}_m - x_m)
\]

\[
\tilde{x}_m \leftarrow \tilde{x}_m - \mu' \frac{df}{dx_m}
\]

In this scheme, the classification result after attacking is conventionally not freely controlled. Alternatively, \( y_l \) can be selected based on the intended category instead of the second largest element. In this case, the scheme is referred to as “selective attack”.

### III. RESULTS AND DISCUSSIONS

The ONN system is first employed to classify digit images (MNIST dataset [30]) and product images (Fashion-MNIST dataset [31]). The number of pixels in each input image is \( M = 28 \times 28 = 784 \) and the number of categories is \( N = 10 \). Therefore the size for the weighting matrix will be \( 10 \times 784 \). The learning parameter is set as \( \mu = 0.0001 \). The number of iterations is 300. Totally 10000 samples are used to train the system and then another 10000 different images are used for testing. A classification accuracy of 91.3% (for MNIST images) and 83.5% (for Fashion-MNIST images) can be achieved.

Then the adversarial attacks described above are performed on 200 random test images of number digits (or fashion products) that are already correctly classified. The parameters are set as \( \alpha = 1, \beta = 1, \gamma = 2 \) and \( \mu' = 0.01 \). The number of iterations is 300. Finally, 84% (for number digits) and 88% (for fashion products) of the attacked images become mis-classified. Some examples of successful attacked results are shown in Fig. 4.

It can be observed that the attacked images are visually very similar to the original ones. They appear to be contaminated with noise very slightly. But it shall be noticed that the “noise” in this task is optimized, in comparison to random noise. The classification results for the original images and the attacked
Fig. 4. Examples of successful adversarial attack results on an ONN with cascaded MZIs for (a) number digit image classification; (b) fashion product image classification (left: image; right: classification result).

images are very likely to be identical from human eyes. However, the actual classification results of the attacked images by using the ONN with cascaded MZIs are all incorrect, different from those of the original images. For two pairs of original images and attacked images, it can be observed that the maximum elements in the output vectors are changed, shown in Fig. 5. The ONN system is “cheated” and “fooled” by most of our attacked input images.

The relationship between attacking performance and attacked image quality is further analyzed. The former can be measured by the mis-classification rate after attacking and the latter can be measured by the average Peak Signal-to-Noise Ratio (PSNR) between original images and attacked images. Ideally, both a good performance with high mis-classification rate and high image fidelity after attacking are favorable. However, it is a trade-off in practice and two requirements cannot be satisfied simultaneously. In Formula (3), the parameters $\alpha$, $\beta$, and $\gamma$ can be tuned to balance these two-evaluation metrics. When $\alpha = 1$, $\beta = 1$, and the value of $\gamma$ is adjusted from 1 to 3, the relationship between attacking success rate and attacked image quality is shown in Figs. 6 and 7 for number digit images and fashion product images respectively. As a comparison, random Gaussian noise with a standard deviation increasing from 0.045 to 0.145 is added to the original images. The attacking success rate and attacked image quality with random Gaussian noise are evaluated as well. From Figs. 6 and 7, it is evident that the misclassification rate of images that are originally correctly recognized by the ONN classifier is much higher in our proposed scheme than that in the random Gaussian noise scheme when the attacked image quality is similar. In other words, a small amount of noise can yield a good attacking performance in our proposed
scheme while a large amount of random Gaussian noise will still not degrade the performance of ONN classifier significantly. As stated above, the added noise is optimized adaptively regarding to a specific machine learning task in our adversarial attack. Therefore, it is much more effective than random noise.

Some examples of successful attacks with our proposed scheme and corresponding failure attacks with random Gaussian noise to the same set of images are shown in Fig. 8. It can be observed that the random noise is more evenly distributed across the entire image. The ONN classifier is inherently robust to a certain level of random noise. The classification result may still remain correct even if a considerable level of random noise is added on the input image. But the “optimized noise” in our proposed adversarial attack scheme is mainly distributed in some critical feature regions and the minor changes it causes may significantly disturb the output result of the ONN image classifier.

The classification results after attack shown above are not pre-defined. As a more challenging task, a “selective” adversarial attack is further implemented. Each image is attacked with a target including all of the remaining nine other categories. The parameters are set as $\alpha = 1$, $\beta = 1$, $\gamma = 1$ and $\mu' = 0.01$. The number of iterations is 300. The successful rate of attack is reduced to 61.87% (for MNIST images) and 50% (for Fashion-MNIST images). Some examples of selective attacking results for MNIST number digit images are shown in Fig. 9.

It can be observed that the original image can be transformed to another image with minor changes that will be mis-classified as the corresponding pre-defined desirable one of the remaining nine categories, despite the failure cases. The differences between original images and attacked images in Fig. 9 are slightly higher than those in Fig. 4.

**IV. Conclusion**

The adversarial attack can substantially degrade the classification performance of an ONN system by generating incorrect output results from apparently similar input images. This phenomenon is significant for both electronic and optical machine learning systems. The investigation of adversarial attacks on linear classifiers based on vector-matrix multiplication also applies.
to other similar optical machine learning systems [32], [33], [34], [35], in addition to an ONN with cascaded MZIs.

In future works, the exploration of adversarial attacks on ONNs will be conducted from the following two aspects. First, adversarial attacks on other ONN models that are more complicated than a linear classifier will be carried out. The mathematical relationship between input samples and output results can be still revealed by a network model consisting of linear and non-linear functions. The minimum perturbations on the test samples for producing significant changes in the output results can be found by proper optimization algorithms. The basic adversarial attack strategies will be similar for both linear and nonlinear ONN models. Second, the robustness of an ONN will be further enhanced against these adversarial attacks. The possible methods for defending an optical neural network against adversarial attacks can be investigated from two perspectives: modification of training or testing samples and improvement of network model. Both can be implemented either optically or digitally. The training dataset may include adversarial examples and the mis-classification performance will be corrected in a straightforward manner. The testing samples may be processed by certain filtering or compression in advance to remove the perturbations used for attacking. The structure, optimization algorithm and loss function of network model can be modified in tailor-made ways against adversarial attacks such as regularization and distillation.

REFERENCES

[1] C. Huang et al., “Prospects and applications of photonic neural networks,” Adv. Phys.: X, vol. 7, no. 1, pp. 1–63, 2022.
[2] J. Liu et al., “Research progress in optical neural networks: Theory, applications and developments,” PhotoniX, vol. 2, no. 1, pp. 1–39, Apr. 2021.
[3] Y. Shen et al., “Deep learning with coherent nanophotonic circuits,” Nat. Photon., vol. 11, no. 7, pp. 441–446, Jun. 2017.
[4] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, “Training of photonic neural networks through in situ backpropagation and gradient measurement,” Optica, vol. 5, no. 7, pp. 864–871, Jul. 2018.
[5] H. Zhang et al., “An optical neural chip for implementing complex-valued neural network,” Nat. Commun., vol. 12, no. 1, pp. 1–11, Jan. 2021.
[6] T. Zhang et al., “Efficient training and design of photonic neural network through neuroevolution,” Opt. Express, vol. 27, no. 26, pp. 37150–37163, Dec. 2019.
[7] H. Zhou et al., “Chip-scale optical matrix computation for PageRank algorithm,” IEEE J. Sel. Topics Quantum Electron., vol. 26, no. 2, Mar./Apr. 2020, Art. no. 8300910.
[8] X. Lin et al., “All-optical machine learning using diffractive deep neural networks,” Science, vol. 361, no. 6406, pp. 1004–1008, Jul. 2018.
[9] J. Shi, Y. Chen, and X. Zhang, “Broad-spectrum diffractive network via ensemble learning,” Opt. Lett., vol. 47, no. 3, pp. 605–608, Feb. 2022.
[10] Y. Xiao, S. Li, G. Sittu, and Z. You, “Optical random phase dropout in a diffractive deep neural network,” Opt. Lett., vol. 46, no. 20, pp. 5260–5263, Oct. 2021.
[11] H. Chen et al., “Diffractive deep neural networks at visible wavelengths,” Engineering, vol. 7, no. 10, pp. 1483–1491, Oct. 2021.
[12] D. Mengü, Y. Liao, Y. Rivenson, and A. Özcan, “Analysis of diffractive optical neural networks and their integration with electronic neural networks,” IEEE J. Sel. Topics Quantum Electron., vol. 26, no. 1, Jan./Feb. 2020, Art. no. 3700114.
[13] X. Xu et al., “11 TOPS photonic convolutional accelerator for optical neural networks,” Nature, vol. 589, no. 7840, pp. 44–51, Jan. 2021.
[14] Y. Yao et al., “All-optical neural network with nonlinear activation functions,” Optica, vol. 6, no. 9, pp. 1132–1137, Sep. 2019.
[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014, arXiv:1412.6572.
[16] N. Akhtar and A. Mian, “Threat of adversarial attacks on deep learning in computer vision: A survey,” IEEE Access, vol. 6, pp. 14410–14430, 2018.
[17] A. Gnanasambandam, A. M. Sherman, and S. H. Chan, “Optical adversarial attack,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 92–101.
[18] Z. Guo et al., “Adversarial attack against deep saliency models powered by non-redundant priors,” IEEE Trans. Image Process., vol. 30, pp. 1973–1988, 2021.
[19] M. Xu, T. Zhang, Z. Li, M. Liu, and D. Zhang, “Towards evaluating the robustness of deep diagnostic models by adversarial attack,” Med. Image Anal., vol. 69, Apr. 2021, Art. no. 101977.
[20] S. Hu et al., “Adhvas: Set-to-target targeted attack on deep hashing with one single adversarial patch,” in Proc. 29th ACM Int. Conf. Multimed., Oct. 2021, pp. 2335–2343.
[21] S. Hu et al., “Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 15014–15023.
[22] A. E. Cià, A. Torcinovich, and M. Pellillo, “A black-box adversarial attack for poisoning clustering,” Pattern Recognit., vol. 122, Feb. 2022, Art. no. 108306.
[23] B. Peng, B. Peng, J. Zhou, J. Xia, and L. Liu, “Speckle-variant attack: Toward transferable adversarial attack to SAR target recognition,” IEEE Geosci. Remote Sens. Lett., vol. 19, 2022, Art. no. 4509805.
[24] V. Kravets, B. Javidi, and A. Stern, “Compressive imaging for defending deep neural networks from adversarial attacks,” Opt. Lett., vol. 46, no. 8, pp. 1951–1954, Apr. 2021.
[25] V. Kravets, B. Javidi, and A. Stern, “Compressive imaging for thwarting adversarial attacks on 3D point cloud classifiers,” Opt. Exp., vol. 29, no. 26, pp. 42726–42737, Dec. 2021.
[26] K. Kim et al., “Engineering pupil function for optical adversarial attacks,” Opt. Exp., vol. 30, no. 5, pp. 6500–6518, Feb. 2022.
[27] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, “Dive into deep learning,” 2021, arXiv:2106.11342.
[28] F. Shokranesh, S. Geoffrey-Gagnon, N. M. Nezami, and O. Liboiron-Ladouceur, “A single layer neural network implemented by a 4 × 4 MZI-Based optical processor,” IEEE Photon. J., vol. 11, no. 6, Dec. 2019, Art. no. 4501612.
[29] F. Shokranesh, N. M. Nezami, and O. Liboiron-Ladouceur, “Theoretical and experimental analysis of a 4 × 4 reconfigurable MZI-based linear optical processor,” J. Lights. Technol., vol. 38, no. 6, pp. 1258–1267, Mar. 2020.
[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-Based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[31] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms,” 2017, arXiv:1708.07747.
[32] S. Jiao et al., “Optical machine learning with incoherent light and a single-pixel detector,” Opt. Lett., vol. 44, no. 21, pp. 5186–5189, Nov. 2019.
[33] P. Léonard, E. J. Fuller, C. M. Teeter, and C. M. Vineyard, “High accuracy single-layer free-space neural classifiers for spatially incoherent light,” Opt. Exp., vol. 30, no. 8, pp. 12510–12520, Apr. 2022.
[34] B. Limbacher et al., “Terahertz optical machine learning for object recognition,” APL Photon., vol. 5, no. 12, Dec. 2020, Art. no. 126103.
[35] L. Li et al., “Machine-learning reprogrammable metasurface imager,” Nat. Commun., vol. 10, no. 1, Mar. 2019, Art. no. 1082.

Shuming Jiao received the Ph.D. degree in electronic engineering from the City University of Hong Kong, Hong Kong, in 2016. He is currently an Assistant Researcher with Peng Cheng Laboratory, Shenzhen, China. His research interests include holographic imaging and display, single-pixel imaging, optical computing, optical security, image processing, and machine learning.

Ziwei Song was born in Jiaozuo, China, in 1996. She is currently working toward the Ph.D. degree with Xidian University, Xi’an, China. Her research interests include photonic neuromorphic computing, photonic spiking neural networks, vertical cavity surface-emitting lasers, and semiconductor lasers.

Shuiliing Xiang was born in Jin’An, China, in 1986. She received the Ph.D. degree from Southwest Jiaotong University, Chengdu, China, in 2013. She is currently an Associate Professor with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi’an, China. She is the author or co-author of more than 80 research papers. Her research interests include photonic neuromorphic computing, brain-inspired information processing, spiking neural network, vertical cavity surface-emitting lasers, and semiconductor lasers dynamics.