# Late Breaking Results: Physical Adversarial Attacks of Diffractive Deep Neural Networks

Yingjie Li*, Cunxi Yu†
Department of Electrical and Computer Engineering
University of Utah
Salt Lake City, UT, USA
Email: *u1306858@utah.edu, †cunxi.yu@utah.edu,

2021 58th ACM/IEEE Design Automation Conference (DAC) | 978-1-6654-3274-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/DAC18074.2021.9586204

***Abstract*—Diffractive Deep Neural Network (D$^2$NN) can work as a neural network with the diffraction of light and have demonstrated orders of magnitude performance improvements in computation speed and energy efficiency [1], [2]. As a result, there have been increasing interests in applying D$^2$NNs into security-sensitive applications, such as security gate sensing, drug detection, etc. However, the comprehensive vulnerability and robustness of optical neural networks have never been studied. In this work, we develop the first adversarial attack formulations over optical physical meanings, and provide comprehensive analysis of adversarial robustness of D$^2$NNs under practical adversarial threats over optical domains, i.e. Phase attack, Amplitude attack, and Complex-domain attack, which can be realized in D$^2$NN system using amplitude and phase modulators. We demonstrate that the proposed Complex Fast Gradient Sign Method (Complex-FGSM) can successfully generate minimal-changed (small epsilon) physically feasible adversarial examples targeting pre-trained D$^2$NNs model on MNIST-10 dataset, which bring down its accuracy to $\leq 20\%$ from 95.4%.**

***Index Terms*—Optical neural networks, security, adversarial learning**

## I. Introduction and Motivation

Nowadays, there have been increasing efforts in leveraging optics to overcome defeats of conventional Neural networks, which will bring significant advantages in power efficiency, parallelism, and computational speed [1]–[4]. Diffractive Deep Neural Networks (D$^2$NNs) utilize the diffraction of light in complex domain to form an optical feed-forward network similar to conventional neural network [1]. The forward function in D$^2$NN is based on free-space light propagation, featuring millions of neurons in each layer interconnected with neurons in neighboring layers, making the system able to complete parallel tasks in the speed of light [1] [2]. Moreover, physical parameters in diffractive propagation are differentiable such that they can be effectively optimized via conventional backpropagation algorithms using `autograd` mechanism [1] [2]. Increasing efforts are devoted to applying such networks in real-world scenarios such as medical sensing, security screening, drug detection, and autonomous driving, which are usually highly sensitive to the security threats [5]. However, very few researches have been conducted in studying comprehensive vulnerability and robustness of neural networks in optical domain (complex tensor domain).

As deep neural networks have been widely used in real-world applications, security and integrity of the applications pose great concern. In some cases, adversaries can be dangerous as it may be imperceptible to human eyes but can force a trained model to produce incorrect outputs [6]. Specifically, with limited exploration of adversarial attack in optical domains under domain-specific physical meanings, the adversarial threats in optical neural networks remain unknown. Thus, this work introduces three attack modes under physical meanings over optical domain, including **a)** Amplitude attack, in which only the real part in perturbation will be applied to the *Real* part of the input, **b)** Phase attack, in which only the *Imaginary* part in perturbation will be applied into the input, **c)** Complex-domain attack, where the perturbation noise is a complex tensor that will be applied to the input data. Note that in D$^2$NNs, the image is encoded using light source, where the original image is encoded using Amplitude (real part), and Phase remains zero (imaginary), such that the inputs are complex tensors. For the attack algorithm, we modify the Fast Gradient Sign Method (FGSM) in complex domain, namely Complex-FGSM (C-FGSM). Moreover, we explore the effectiveness of C-FGSM by adjusting the hyperparameter epsilon that describes how much the original data is modified. This is believed to be the first work on adversarial robustness of optical neural networks under physical meanings.

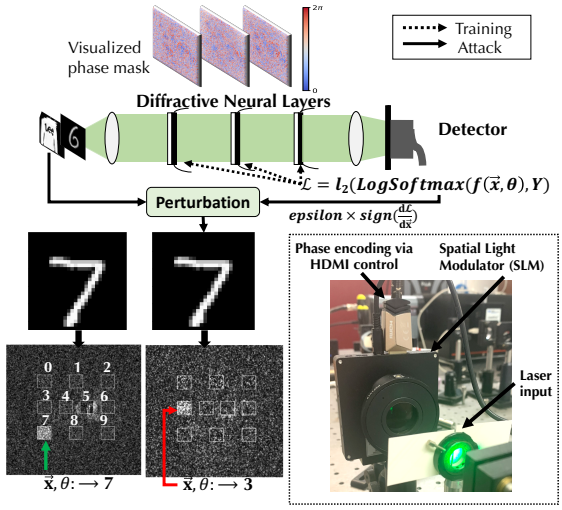


Fig. 1: **Illustration of D$^2$NNs system, training, and adversarial attack using C-FGSM** – Note that in training phase, the optical phase encoded by phase modulators are trainable parameters $\theta$ optimized through gradient descent. The adversarial examples are generated with C-FGSM according to the loss function $\mathcal{L}$. An inference example with original MNIST image (bottom left) shows input and output observed by detector, where class `7` is generated. An adversarial example (bottom right) is however misclassified to `3`.

## II. Methods and Results

Our experiment setup is shown in Figure 1. The original MNIST images with size of $28 \times 28$ are first expanded to $200 \times 200$ to fit our Spatial Light Modulator (SLM) based optical system setup. Three layers are implemented in the model. Each layer is composed of a diffractive layer and a phase modulator. The diffractive layer is used to diffract light so that each pixel in the layer can work as a neuron in conventional neural network, i.e., each diffraction layer mimics one conventional linear neural layer. The phase information encoded by

the phase modulator is a trainable parameter in our model and can be configured by the voltage applied to each pixel of the modulator (HDMI control shown in Figure 1). The forward function is described as follows:

$$\begin{aligned} \mathcal{F} &= LogSoftmax(det(f(\vec{x}, \theta)) \\ f &: \mathcal{X} \rightarrow (\Re + i\Im)^{\in 200 \times 200} \end{aligned} \tag{1}$$

where the function $f$ is the map for the model, the vector $\vec{x}$ is the original input image whose size is $200 \times 200$ in complex domain with the imaginary part initialized as all zeros for original MNIST data, where $\Re$ represents the real part and $\Im$ represents the imaginary part in the complex domain. The $D^2NNs$ parameters $\theta = \{\theta_0, \theta_1, \theta_2\}$ describe the encoded phase parameters of the three diffractive layers with the size of $200 \times 200$. The detector $det$ is used to capture the output of the diffractive layers, which will be used to produce the prediction result $\mathcal{F}$. There are 10 separate regions on the detector representing label 0 to 9 as shown in Figure 1. The loss function is implemented with `L2-norm`.

The adversarial perturbations are generated by C-FGSM with hyperparameter *epsilon* which is used to control the size of the perturbation noise w.r.t the original images. Considering the physical meaning in optical systems, clamp functions are needed to make sure adversarial examples are physically feasible. First, in an optical system, we use laser as the light source, which requires all pixels in the input image to be bounded to a certain amplitude range. Thus, the real part will be clamped into [0, 1] after normalization. Second, since the working range of the Spatial Light Modulator (SLM) in our system is [0, $\pi$], the imaginary part will be clamped into [0, $\pi$]. Future works will focus on adversarial countermeasures and adversarial example fabrications.

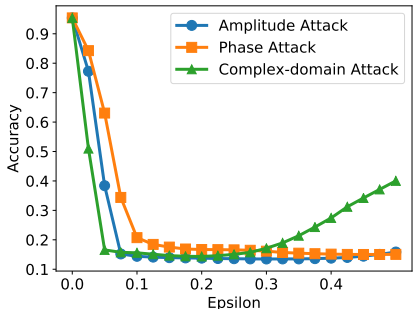


Fig. 2: **Evaluation of proposed C-FGSM on attacking pre-trained $D^2NNs$ (acc=95.4%) using MNIST-10 dataset.**

TABLE I: Accuracy under three attack modes with large epsilons.

| Epsilon | Phase attack | Amplitude attack | Complex attack |
|---|---|---|---|
| 1.0 | 28.0% | 50.4% | 60.8% |
| 2.0 | 51.3% | 50.4% | 63.0% |
| 3.0 | 57.5% | 50.4% | 63.6% |
| 4.0 | 58.1% | 50.4% | 63.7% |

Figure 2 shows the effectiveness of our attack with C-FGSM under different attack modes. As we can see from Figure 2, for all three attack modes, the model can be attacked efficiently when epsilon is $\leq 0.2$. For example, when the epsilon is 0.1, the attacks will degrade the accuracy by 75% from 95.4% under all three attack modes. The most effective attack mode is Complex-domain attack and the least effective mode is Phase attack. For Complex-domain attack, it will attack both real and imaginary part, which decreases the accuracy very efficiently. For Phase attack, since the phase for all original input image are identical (all zeros), it is not an 'authentic' feature in objects. The 'featured' phase for all images are generated from the model. As a result, the attack posed on phase will be the least effective. However, when the epsilon is large enough ($\geq 0.25$), the accuracy under Complex-domain attack will increase as the epsilon increases. The similar observation will be found when epsilon is larger than 0.3 under Amplitude attack and when epsilon is larger than 0.5 under Phase attack.

The observations are mainly caused by the clamping and formulations of different attack modes. In Complex-domain attack, we apply both real part and imaginary part perturbation to the original data. While in Amplitude and Phase attack, the corresponding perturbation part only consists of either amplitude or phase perturbation. As a result, Complex-domain adversarial example includes twice of the perturbation than other two attacks, i.e., the input images under Complex-domain attack will be perturbed and clamped twice, for real part and imaginary part separately. When epsilon is too large ($\geq 0.25$), more data in the original image will be clamped, which means more parts of our attack with the sign of the gradient will be ineffective (Table I). As a result, our attack with gradient sign intended to decrease the accuracy will be less powerful. Table I included the accuracy of the model under different attacks with large epsilon ($\geq 1.0$). The accuracy for Complex-domain attack will always stay the same (63.7%) when epsilon is larger than 3.14 ($\pi$); the accuracy under Amplitude attack will stay the same (50.4%) when epsilon is larger than 1; the accuracy under Phase attack will stay the same (58.1%) after 3.14. This is because when epsilon is greater than 1.0 or $\pi$, the adversarial example of the real part will binarized real tensor (0 or 1) or "binarized" phase tensor (0 or $\pi$), respectively. Then, the adversarial image dataset after perturbation and clamp will remain the same no matter how much epsilon increases.

To implement our attack in real-world application, we can apply some defeats (gray points) according to the perturbation onto the input image to realize Amplitude attack, apply a thin transparent film designed with different thickness to modify the phase of input images according to the perturbation on the input image to realize Phase attack. For Complex-domain attack, we will combine defeats and transparent film together to create the adversarial image. Since our effective epsilon is very small ($\leq 0.2$), the defeats and film will be imperceptible to human eyes but it can confuse the model and degrade the accuracy efficiently.

**Acknowledgment** This work is supported by U.S. National Science Foundation (NSF) awards NSF-2019336 and NSF-2008144.

## REFERENCES

[1] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, 2018.

[2] D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Analysis of diffractive optical neural networks and their integration with electronic neural networks," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–14, 2019.

[3] J. Gu, Z. Zhao, C. Feng, W. Li, R. T. Chen, and D. Z. Pan, "Flops: Efficient on-chip learning for optical neural networks through stochastic zeroth-order optimization," in *(DAC'2020)*. IEEE, 2020, pp. 1–6.

[4] J. Gu, Z. Zhao, C. Feng, M. Liu, R. T. Chen, and D. Z. Pan, "Towards area-efficient optical neural networks: an fft-based architecture," in *ASP-DAC'2020*. IEEE, 2020, pp. 476–481.

[5] J. F. Federici, B. Schulkin, F. Huang, D. Gary, R. Barat, F. Oliveira, and D. Zimdars, "Thz imaging and sensing for security applications—explosives, weapons and drugs," *Semiconductor Science and Technology*, vol. 20, no. 7, p. S266, 2005.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.