Deep Learning Based Attack on Phase-Truncated Optical Encoding

The following publication L. Zhou, X. Chen and W. Chen, "Deep Learning Based Attack on Phase-Truncated Optical Encoding," 2020 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO), Hangzhou, China, 2020 is available at https://doi.org/10.1109/NEMO49486.2020.9343452.

LINA ZHOU
Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong, China.
The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518057, China

XUDONG CHEN
Department of Electrical and Computer Engineering,
National University of Singapore,
Singapore 117583, Singapore

WEN CHEN
Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University,
Hong Kong, China
owen.chen@polyu.edu.hk

***Abstract*—We apply the learning based attack to study the vulnerability of phase-truncated optical encoding scheme. By using a number of ciphertext-plaintext pairs to train a designed learning model, an attacker can effectively analyze the vulnerability of optical encryption scheme based on phase truncation. The learning based attacks for phase-truncated optical encoding can retrieve unknown plaintexts from the given ciphertexts, which can avoid the retrieval of security keys and the design of complex phase retrieval algorithms. It is demonstrated that the learning based attack can provide a promising approach for vulnerability analysis of phase-truncated optical cryptosystems.**

Keywords—learning based attacks, optical encoding, phase truncation

I. INTRODUCTION

In recent years, optical technology is of increasing interest for many applications in optical cryptography [1–19]. The application of optical technologies to optical encryption can be ascribed to their intrinsic and extraordinary properties, e.g., parallel processing and multi-dimensional processing. In the early stage, it was demonstrated that an image (i.e., plaintext) can be encrypted by using one phase-only mask placed at the input image plane and another phase-only mask placed at Fourier domain [1]. This scheme known as double random phase encoding (DRPE) was proposed as the basis for optical encryption [1]. By applying the DRPE scheme, the plaintext was encoded into a stationary noise pattern (i.e., the ciphertext). Research had been widely carried out to implement DRPE in various domains, e.g., fractional Fourier transform and Fresnel transform [2–8]. As validated by many researchers, the DRPE approach has advantages of high feasibility and high flexibility in practical applications. Other optical techniques were also found to be effective for optical encryption, e.g., diffractive imaging, interferometry and ghost imaging [9–13]. Many optical encryption techniques have been developed for the cryptography until now, therefore vulnerability of the optical security systems needs to be further analyzed.

For optical cryptography, cryptoanalysis is usually not carried out in detail. In essence, there is a mutual benefit in the relationship between cryptography and cryptoanalysis. It is expected that cryptographic techniques are secure enough to withdraw the attacks by cryptoanalysis. Meanwhile, the techniques for the cryptoanalysis further promote the development of more advanced cryptographic methods. Some research has been done to develop the cryptoanalysis [20–27]. Vulnerability analysis of the DRPE scheme by using chosen-ciphertext attack (CCA) has been first proposed [20]. Then, chosen-plaintext attack has also been developed [21]. In addition, ciphertext-only attack and known-plaintext attack have been designed and proven to be effective for retrieving unknown plaintexts from the ciphertexts [22–24]. The forementioned cryptoanalysis techniques rely on the retrieval of security keys and the usage of complex phase retrieval algorithms, which can hinder the wider applications of cryptoanalysis owing to the difficulty in practical applications. Recently, a method called learning based attack was developed, and was verified to be effective for attacking diffraction-imaging-based encryption, optical-interference-based encryption and CGH-based encryption [25–27]. It is different from conventional methods, since it is free from the extraction of security keys and the usage of complex phase retrieval algorithms. In the learning based attacks, a designed learning model is trained by using ciphertext-plaintext pairs, and then the trained learning model can retrieve unknown plaintext from the ciphertext without information about the security keys and without the usage of complex phase retrieval algorithms [25–27]. Although security keys are not directly retrieved, the trained learning model is able to successfully retrieve transfer function of the optical security systems.

In this paper, we apply the learning based attacks to study the vulnerability of phase-truncated optical encoding scheme. The plaintexts are encrypted by using phase-truncated optical encryption scheme [7,28,29]. Using the pairs of ciphertexts and plaintexts to train a designed learning model, the trained learning model can predict unknown plaintext from the ciphertext without encryption keys. It is demonstrated that the learning based attacks are effective, and can be applied to analyze the security or vulnerability of phase-truncated optical encoding scheme.

II. THEORY

Figure 1(a) shows a schematic setup for the phase-truncated optical cryptosystem. The procedure of phase truncation-based optical encoding is as follows: the first phase-only mask M_1 [i.e., $M_1(x,y)$] is placed just behind an input image (i.e., the plaintext used in this study), and the wave propagation in Fourier domain can be described by

$$w_1(\mu,\upsilon) = FT[p(x,y)M_1(x,y)], \tag{1}$$

where $w_1(\mu,\upsilon)$ denotes a wave pattern in Fourier domain, $p(x,y)$ denotes the plaintext, and FT denotes free-space wave propagation in Fourier domain. Then, phase information of $w_1(\mu,\upsilon)$ is truncated, and amplitude information of $w_1(\mu,\upsilon)$ is reserved for the further encoding. The reserved amplitude information can be described by

$$Am(\mu,\upsilon) = |w_1(\mu,\upsilon)|. \quad (2)$$

The amplitude-only pattern $Am(\mu,\upsilon)$ locates just before the second mask M_2 [i.e., $M_2(\mu,\upsilon)$], and wave propagation is further implemented and can be described by

$$w_2(\xi,\eta) = IFT[Am(\mu,\upsilon)M_2(\mu,\upsilon)], \quad (3)$$

where *IFT* denotes inverse Fourier transform. Similarly, phase truncation is also implemented, and the reserved amplitude pattern can be described by

$$c(\xi,\eta) = |w_2(\xi,\eta)|, \quad (4)$$

where $c(\xi,\eta)$ denotes the ciphertext obtained by using the phase truncation-based optical cryptosystem. In phase-truncated optical encoding, the truncated phase pattern at the mask M_2 plane and the CCD plane are considered as keys. It should be emphasized that phase-mask keys change when the input plaintexts are changed [7]. Hence, it has been claimed in previous studies that phase-truncated optical encoding is secure, which can withdraw the attacks. Here, we find that the learning based attacks can also be applied to analyze the vulnerability of phase-truncated optical encoding scheme. The machine learning based attacks do not rely on the direct retrieval of different security keys. By using a trained learning model, the unknown plaintexts can be effectively retrieved from the given ciphertexts in real time.

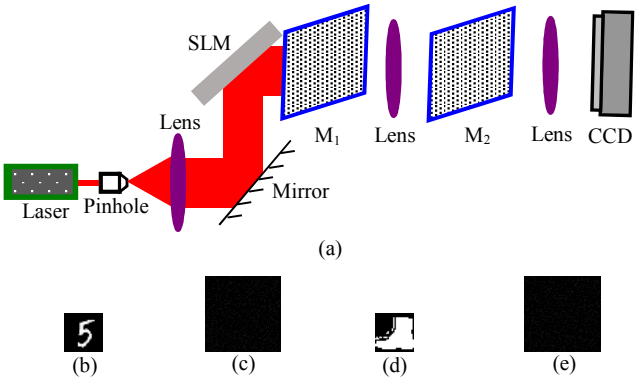

Fig. 1. (a) A schematic setup for phase-truncated optical encoding scheme. Phase truncation is conducted at the M2 plane and the CCD plane. SLM: spatial light modulator; M1: the first phase-only mask; M2: the second phase-only mask; CCD: charge-coupled device. (b) A handwritten-digit image from MNIST database and (d) a fashion-product image from fashion MNIST database. (c) and (e) The ciphertexts respectively corresponding to (b) and (d).

In the phase-truncated optical cryptosystem shown in Fig. 1(a), after the phase truncation is implemented in the M_2 plane, the wave further propagates through M_2 and then is recorded. The inputs (i.e., plaintext) are handwritten-digit images from MNIST database and fashion-product images from fashion MNIST database [30,31]. Using a series of input images from these two databases, the corresponding ciphertexts are obtained by using the optical encoding scheme. Figures 1(b) and 1(c) show a handwritten-digit image from the MNIST database and its corresponding ciphertext obtained by using the phase-truncated optical cryptosystem, respectively. An example of a fashion-product image from the fashion MNIST database and its corresponding ciphertext are shown in Figs. 1(d) and 1(e).

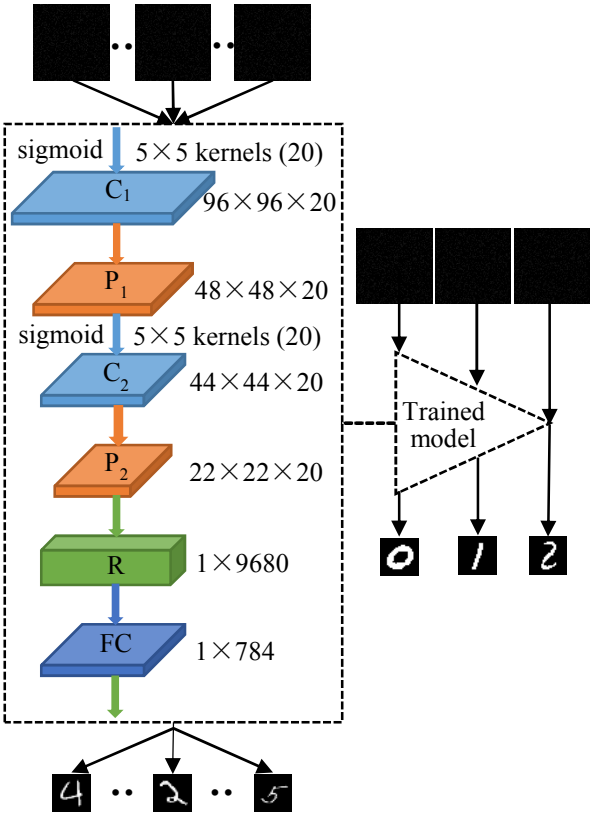

Fig. 2. Schematic of the CNN model for attacking phase-truncated optical encoding scheme: C_1: the first convolutional layer; P_1: the first pooling layer; C_2: the second convolutional layer; P_2: the second pooling layer; R: reshaping layer; FC: Fully connected layer. The inputs are the ciphertexts obtained by using the optical cryptosystem, and the outputs are the corresponding plaintexts. The designed CNN model contains two convolutional layers, two pooling layers, one reshaping layer and one fully connected layer. After the training, the CNN model can be applied to retrieve unknown plaintexts in the phase-truncated optical cryptosystem.

Machine learning is widely applied nowadays to recover information from a speckle pattern [32,33]. Extensive applications of machine learning have also pushed the development of the learning based attacks. Machine learning based attack is a new method, which has been verified to be applicable to analyze the security of optical encryption [25–27]. Here, a deep learning model [27] is applied for the vulnerability analysis of phase-truncated optical encoding scheme. The designed learning model, called convolutional neural network (CNN), is illustrated in Fig. 2. The intensity patterns (i.e., ciphertexts) obtained by the phase-truncated optical cryptosystem and their corresponding input images (i.e., the plaintexts) are fed to the designed CNN model respectively as inputs and outputs. The ciphertexts with 100×100 pixels are resized from intensity patterns with 512×512 pixels. The plaintexts with 28×28 pixels are from the MNIST database. Architecture of the designed CNN model is as follows: The ciphertext (100×100 pixels) is convolved with 20 kernels with a dimension of 5×5, and the first convolutional layer is generated with a size of 96×96×20. Activation function used in the first convolution process is a sigmoid function. Then, the first convolutional layer is transferred to the first pooling layer (size of 48×48×20) under the processing of down sampling to lower the computational load.

Subsequently, the first pooling layer convolves 20 kernels (5×5) forming the second convolutional layer (44×44×20). The sigmoid function is adopted again in the process of convolution. To further reduce the dimension of the second convolutional layer, down-sampling is implemented in the second convolutional layer, which forms the second pooling layer (22×22×20). Since the second pooling layer is three-dimensional, huge computational load is requested to conduct vector calculations. Hence, the second pooling layer is reshaped to one-dimensional vector (1×9680). To correlate with the corresponding plaintext, an operation of fully connection is applied following the reshaped layer, which forms the fully connected layer with a size of 1×784. After reshaping, the fully connected layer is reshaped to a two-dimensional vector (28×28). The obtained two-dimensional vector is the predicted image given by the designed CNN model. Mean squared error (MSE) is employed to evaluate the difference between the prediction and original plaintext. When the MSE is higher than a preset value, the errors between the predicted image and original plaintext are back-propagated to update the parameters used for designing the CNN model. The updating rule is stochastic gradient descent, which is the same as that in Ref. 27. The learning rate is 10^{-6}, and the initial value of weights and biases are set to zero. The value of momentum is -0.00095. A series of ciphertexts obtained by the phase-truncated optical encoding scheme and their corresponding plaintexts are sequentially inputted into the designed CNN model, and the parameters of the designed CNN model are continuously updated. Parameters of the CNN model are optimized, which can be utilized to estimate unknown plaintext from the given ciphertexts. Several examples for the testing are shown in Fig. 2. The CNN model is conducted by using a computer with Matlab2009 platform, a Nvidia Geforce GTX1080Ti GPU and RAM of 64G.

III. RESULTS AND DISCUSSION

The 5000 handwritten-digit images from MNIST database and 5000 fashion-product images from fashion MNIST database are used as the plaintexts. Hence, 10000 ciphertexts, i.e., intensity patterns, are correspondingly obtained by using the phase-truncated optical encryption scheme. For each database, 4800 ciphertext-plaintext pairs are selected to train the designed learning model. Another 200 ciphertext-plaintext pairs are used to test feasibility of the trained CNN model. The total training time is about 4 hours for each database. Figure 3 shows the retrieved plaintexts by using the trained CNN model. It can be seen in Fig. 3 that the ciphertexts shown in the first row cannot visually render any information. By employing the trained CNN model, the unknown plaintexts are effectively extracted without usage of security keys as shown in Fig. 3. Correlation coefficient (CC) is calculated to evaluate the quality. The CC values for Figs. 3(h), 3(i), 3(j), 3(k), 3(l), 3(m) and 3(n) are 0.89, 0.79, 0.82, 0.85, 0.87, 0.84 and 0.68, respectively. The ciphertexts shown in Fig. 4 are obtained by using the fashion MNIST database, and their corresponding plaintexts are retrieved by the trained CNN model. It is found that the retrieved plaintexts are of high quality with CC values of 0.89, 0.83, 0.95, 0.88, 0.94, 0.66 and 0.64 for Figs. 4(h), 4(i), 4(j), 4(k), 4(l), 4(m) and 4(n), respectively. It has been previously claimed that the plaintexts encrypted by using phase-truncated optical encoding scheme can be retrieved only when all correct security keys are used. Here, it is demonstrated that the trained CNN models can successfully retrieve unknown plaintexts without the usage of security keys.

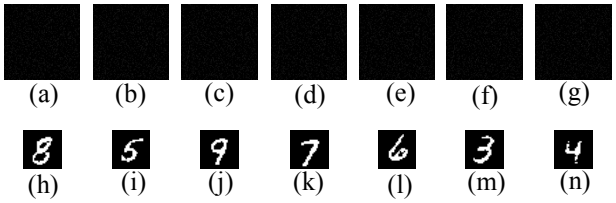

Fig. 3. Testing results using the handwritten-digit MNIST database: (a), (b), (c), (d), (e), (f) and (g) Ciphertexts obtained by using phase-truncated optical encoding. (h), (i), (j), (k), (l), (m) and (n) The plaintexts retrieved by using the trained CNN model.

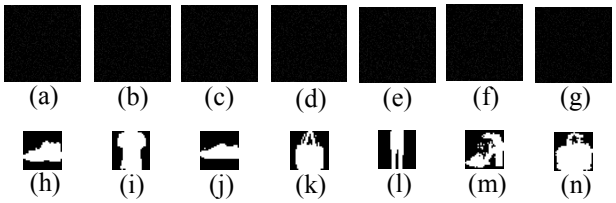

Fig. 4. Testing results using the fashion MNIST database: (a), (b), (c), (d), (e), (f) and (g) Ciphertexts obtained by phase-truncated optical encoding. (h), (i), (j), (k), (l), (m) and (n) The plaintexts retrieved by using the trained CNN model.

IV. CONCLUSIONS

It has been demonstrated that phase-truncated optical encryption scheme cannot withdraw the learning based attacks. Different from conventional cryptoanalysis, the learning based attacks do not request the encryption keys for the plaintext retrieval from the given ciphertexts. It can be expected that the learning based attacks could provide a promising approach for the cryptoanalysis of phase-truncated optical encoding scheme.

ACKNOWLEDGMENTS

The supports from National Natural Science Foundation of China (NSFC) (61605165), Shenzhen Science and Technology Innovation Commission (JCYJ20160531184426473), and Hong Kong Research Grants Council (25201416, C5011-19G) are acknowledged. Xudong Chen acknowledged the support by the National Research Foundation, Prime Minister's Office, Singapore under its Competitive Research Program (CRP Award No. NRF-CRP15-2015-03).

REFERENCES

[1] P. Refregier and B. Javidi, "Optical image encryption based on input plane and Fourier plane random encoding," Opt. Lett., vol. 20, pp. 767–769, April 1995.

[2] N. Singh and A. Sinha, "Gyrator transform-based optical image encryption, using chaos," Opt. Lasers Eng., vol. 47, pp. 539–546, May 2009.

[3] G. Unnikrishnan, J. Joseph, and K. Singh, "Optical encryption by double-random phase encoding in the fractional Fourier domain," Opt. Lett., vol. 25, pp. 887–889, June 2000.

[4] G. Situ and J. Zhang, "Double random-phase encoding in the Fresnel domain," Opt. Lett., vol. 29, pp. 1584–1586, July 2004.

[5] R. Tao, Y. Xin, and Y. Wang, "Double image encryption based on random phase encoding in the fractional Fourier domain," Opt. Express, vol. 15, pp. 16067–16079, November 2007.

[6] L. Chen and D. Zhao, "Optical image encryption with Hartley transforms," Opt. Lett., vol. 31, pp. 3438–3440, December 2006.

[7] W. Chen, B. Javidi, and X. Chen, "Advances in optical security systems," Adv. Opt. Photon., vol. 6, pp. 120–155, April 2014.

[8] O. Matoba and B. Javidi, "Encrypted optical memory system using three-dimensional keys in the Fresnel domain," Opt. Lett., vol. 24, pp. 762–764, June 1999.

[9] W. Chen, X. Chen, and Colin J. R. Sheppard, "Optical image encryption based on diffractive imaging," Opt. Lett., vol. 35, pp. 3817–3819, November 2010.

[10] Y. Shi, T. Li, Y. Wang, Q. Gao, S. Zhang, and H. Li, "Optical image encryption via ptychography," Opt. Lett., vol. 38, pp. 1425–1427, May 2013.

[11] P. Clemente, V. Durán, V. Torres-Company, E. Tajahuerce, and J. Lancis, "Optical encryption based on computational ghost imaging," Opt. Lett., vol. 35, pp. 2391–2393, July 2010.

[12] E. Tajahuerce, O. Matoba, S. C. Verrall, and B. Javidi, "Optoelectronic information encryption with phase-shifting interferometry," Appl. Opt., vol. 39, pp. 2313–2320, May 2000.

[13] W. Chen and X. Chen, "Space-based optical image encryption," Opt. Express, vol. 18, pp. 27095–27104, December 2010.

[14] P. C. Mogensen and J. Glückstad, "A phase-based optical encryption system with polarisation encoding," Opt. Commun., vol. 173, pp. 177–183, January 2000.

[15] A. Alfalou and C. Brosseau, "Optical image compression and encryption methods," Adv. Opt. Photon., vol. 1, pp. 589–636, October 2009.

[16] Y. Zhang, C. H. Zheng, and N. Tanno, "Optical encryption based on iterative fractional Fourier transform," Opt. Commun., vol. 202, pp. 277–285, February 2002.

[17] J. F. Barrera, A. Mira, and R. Torroba, "Optical encryption and QR codes: secure and noise-free information retrieval," Opt. Express, vol. 21, pp. 5373–5378, February 2013.

[18] M. R. Abuturab, "Color image security system based on discrete Hartley transform in gyrator transform domain," Opt. Lasers Eng., vol. 51, pp. 317–324, March 2013.

[19] N. Singh and A. Sinha, "Chaos based multiple image encryption using multiple canonical transforms," Opt. Laser Technol., vol. 42, pp. 724–731, July 2010.

[20] A. Carnicer, M. Montes-Usategui, S. Arcos, and I. Juvells, "Vulnerability to chosen-cyphertext attacks of optical encryption schemes based on double random phase keys," Opt. Lett., vol. 30, pp. 1644–1646, July 2005.

[21] X. Peng, H. Wei, and P. Zhang, "Chosen-plaintext attack on lensless double-random phase encoding in the Fresnel domain," Opt. Lett., vol. 31, pp. 3261–3263, November 2006.

[22] M. Liao, W. He, D. Lu, and X. Peng, "Ciphertext-only attack on optical cryptosystem with spatially incoherent illumination: from the view of imaging through scattering medium," Sci. Rep., vol. 7, pp. 41789, January 2017.

[23] X. Liu, J. Wu, W. He, M. Liao, C. Zhang, and X. Peng, "Vulnerability to ciphertext-only attack of optical encryption scheme based on double random phase encoding," Opt. Express, vol. 23, pp. 18955–18968, July 2015.

[24] X. Peng, P. Zhang, H. Wei, and B. Yu, "Known-plaintext attack on optical encryption based on double random phase keys," Opt. Lett., vol. 31, pp. 1044–1046, April 2006.

[25] L.N. Zhou, Y. Xiao, and W. Chen, "Machine-learning attacks on interference-based optical encryption: experimental demonstration," Opt. Express, vol. 27, pp. 26143–26154, September 2019.

[26] L.N. Zhou, Y. Xiao, and W. Chen, "Vulnerability to machine learning attacks of optical encryption based on diffractive imaging," Opt. Lasers Eng., vol. 125, pp. 105858, February 2020.

[27] L.N. Zhou, Y. Xiao, and W. Chen, "Learning-based attacks for detecting the vulnerability of computer-generated hologram based optical encryption," Opt. Express, vol. 28, pp. 2499–2510, January 2020.

[28] W. Qin and X. Peng, "Asymmetric cryptosystem based on phase-truncated Fourier transforms," Opt. Letters, vol. 35, pp. 118-120, January 2010.

[29] Y. Wang, C. Quan, and C. J. Tay, "Optical color image encryption without information disclosure using phase-truncated Fresnel transform and a random amplitude mask," Opt. Commun., vol. 344, pp. 147–155, June 2015.

[30] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," IEEE Signal Process. Mag., vol. 29, pp. 141–142, November 2012.

[31] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, September 2017.

[32] L.N. Zhou, Y. Xiao, and W. Chen, "Imaging through turbid media with vague concentrations based on cosine similarity and convolutional neural network," IEEE Photon. J., vol. 11, pp. 7801315, August 2019.

[33] L.N. Zhou, Y. Xiao, and W. Chen, "Image recovery through turbid water under wide distance ranges," The 7th International Conference on Optical and Photonic Engineering (icOPEN 2019), vol. 11205, pp. 112051D, October 2019.