

On the Security of Pixel-Based Image Encryption for Privacy-Preserving Deep Neural Networks

Warit Sirichotedumrong, Yuma Kinoshita and Hitoshi Kiya
Tokyo Metropolitan University, Asahigaoka, Hino-shi, Tokyo, 191-0065, Japan

Abstract—This paper aims to evaluate the safety of a pixel-based image encryption method, which has been proposed to apply images with no visual information to deep neural networks (DNN), in terms of robustness against ciphertext-only attacks (COA). In addition, we propose a novel DNN-based COA that aims to reconstruct the visual information of encrypted images. The effectiveness of the proposed attack is evaluated under two encryption key conditions: same encryption key, and different encryption keys. The results show that the proposed attack can recover the visual information of the encrypted images if images are encrypted under same encryption key. Otherwise, the pixel-based image encryption method has robustness against COA.

I. INTRODUCTION

The spread of deep neural networks (DNNs) has greatly contributed to solving complex tasks for many applications [1], [2], such as for computer vision, biomedical systems, and information technology. Deep learning utilizes a large amount of data to extract representations of relevant features, so the performance is significantly improved [3], [4]. However, there are security issues when using deep learning in cloud environments to train and test data, such as data privacy, data leakage, and unauthorized data access. Therefore, privacy-preserving DNNs have become an urgent challenge.

Various methods have been proposed for privacy-preserving computation. The methods are classified into two types: perceptual encryption-based [5]–[16] and homomorphic encryption (HE)-based [17]–[25]. HE-based methods are the most secure options for privacy preserving computation, but they are applied to only limited DNNs [21]–[25]. Therefore, the HE-based type does not support state-of-the-art DNNs yet. Moreover, data augmentation has to be done before encryption. In contrast, perceptual encryption-based methods have been seeking a trade-off in security to enable other requirements, such as a low processing demand, bitstream compliance, and signal processing in the encrypted domain [5]–[16]. A few methods were applied to machine learning algorithms in previous works [5], [6]. The first encryption method [9]–[14] to be proposed for encryption-then-compression (EtC) systems, was demonstrated to be applicable to traditional machine learning algorithms, such as support vector machine (SVM) [5]. However, the block-based encryption method has never been applied to DNNs.

Although a block-based encryption method [6] was applied to image classification with DNNs, in which an adaption

This work was partially supported by Grant-in-Aid for Scientific Research(B), No.17H03267, from the Japan Society for the Promotion Science.

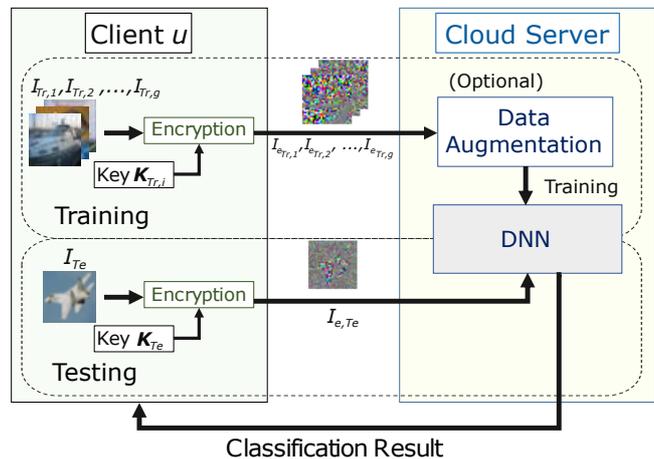


Fig. 1: Scenario

network is added prior to DNNs to avoid the influence of image encryption, the classification performance is inadequate. A pixel-based image encryption method [26] was proposed not only to improve the classification performance of the privacy-preserving DNNs but also to consider data augmentation in the encrypted domain. However, the security level of these encryption methods is only evaluated in terms of the key space analysis for brute-force attack.

In this paper, we aim to evaluate the safety of the pixel-based image encryption in terms of robustness against ciphertext-only attacks (COA), including a new attack called DNN-based ciphertext-only attack. Moreover, the effectiveness of the proposed attack is evaluated under two encryption key conditions: same encryption key, and different encryption keys.

II. SECURITY EVALUATION OF PIXEL-BASED IMAGE ENCRYPTION

A. Privacy-Preserving Deep Neural Networks

Figure 1 illustrates the scenario for privacy-preserving DNNs used in this paper. In the training process, a client u encrypts each training image, $I_{Tr,i}$, $i = 1, 2, \dots, g$, by using a secret key set for training data, $K_{Tr,i}$, and sends the encrypted images ($I_{eTr,i}$) to a cloud server.

In the testing process, the client u encrypts a testing image (I_{Te}) by using a secret key set for testing data, K_{Te} , and sends the encrypted image I_{eTe} to the server. The server solves a classification problem with an image classification model

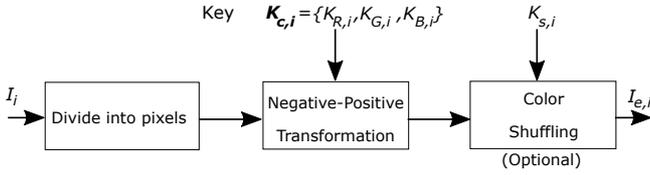


Fig. 2: Pixel-based image encryption

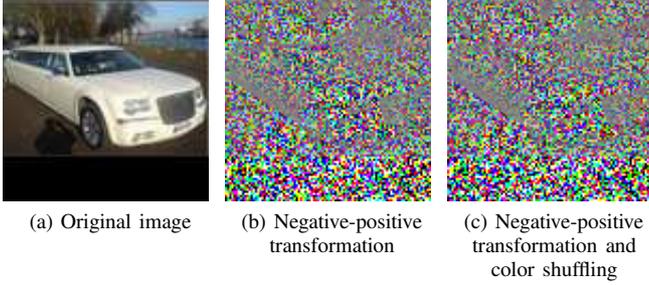


Fig. 3: Examples of images encrypted by proposed method

trained in advance, and then returns the classification results to the client.

In this paper, we assume that there are two encryption key conditions for generating encrypted images as follows.

- **Same encryption key:** All training and testing images are encrypted by using only one secret key, i.e. $K_{Tr,1} = K_{Tr,2} = \dots = K_{Tr,g} = K_{Te} = K$.
- **Different encryption keys:** The different secret keys are randomly assigned to training and testing images, i.e. $K_{Tr,1} \neq K_{Tr,2} \neq \dots \neq K_{Tr,g} \neq K_{Te}$.

B. Pixel-Based Image Encryption

Figure 2 illustrates the encryption steps of the pixel-based image encryption [26]. To generate an encrypted image ($I_{e,i}$) from a color image, I_i , the following steps are carried out, as shown in Fig. 2. Note that the color shuffling (Step 3) is an optional encryption step to enhance security.

- 1) Divide I_i with $X \times Y$ pixels into pixels.
- 2) Individually apply negative-positive transformation to each pixel of each color channel, $I_{R,i}$, $I_{G,i}$, and $I_{B,i}$, by using a random binary integer generated by secret keys $K_{c,i} = \{K_{R,i}, K_{G,i}, K_{B,i}\}$. In this step, a transformed pixel value of the j -th pixel, p' , is calculated using

$$p' = \begin{cases} p & (r(j) = 0) \\ p \oplus (2^L - 1) & (r(j) = 1) \end{cases}, \quad (1)$$

where $r(j)$ is a random binary integer generated by $K_{c,i}$. p is the pixel value of the original image with L bit per pixel. The value of the occurrence probability $P(r(j)) = 0.5$ is used to invert bits randomly [13].

- 3) (Optional) Shuffle three color components of each pixel by using an integer randomly selected from six integers generated by a key $K_{s,i}$ as shown in Table I.

TABLE I: Permutation of color components for random integer. For example, if random integer is equal to 2, red component is replaced by green one, and green component is replaced by red one while blue component is not replaced.

Random Integer	Three Color Channels		
	R	G	B
0	R	G	B
1	R	B	G
2	G	R	B
3	G	B	R
4	B	R	G
5	B	G	R

Images encrypted by using the pixel-based method are illustrated in Fig. 3(b) and 3(c), where Fig. 3(a) is the original one.

III. ROBUSTNESS AGAINST CIPHERTEXT-ONLY ATTACKS

Security mostly refers to protection from adversarial forces. Various attacking strategies, such as the known-plaintext attack (KPA) and chosen-plaintext attack (CPA), should be considered [11]–[13]. In this paper, we consider brute-force attacks and propose a novel DNN-based ciphertext-only attack as ciphertext-only attacks (COA).

A. Brute-force Attack

If a color image I_{RGB} with $X \times Y$ pixels is divided into pixels, the number of pixels n is given by

$$n = X \times Y. \quad (2)$$

The key spaces of negative-positive transformation (N_{np}) and color component shuffling (N_{col}) are represented by

$$N_{np}(n) = 2^{3n}, N_{col}(n) = ({}_3P_3)^n = 6^n. \quad (3)$$

Consequently, the key space of images encrypted by using the proposed encryption scheme, $N(n)$, is represented by the following.

$$N(n) = N_{np}(n) \cdot N_{col}(n) = 2^{3n} \cdot 6^n \quad (4)$$

B. DNN-based Ciphertext-only Attack

We propose a novel DNN-based COA that aims to reconstruct the visual information of encrypted images. Since the encryption method is a pixel-based one, the proposed DNN for COA consists of three 1×1 -locally connected layers, which work similarly to 1×1 -convolution layer, except that weights are unshared. Figure 4 illustrates the proposed attack where $C_j^{M_j}$ is the j -th locally connected layer of the network with a kernel size and stride of $(1,1)$, M_j is the number of feature maps of the j -th locally connected layer, $j \in \{1, 2, 3\}$, and I'_i denotes a reconstructed image. The representations of each encrypted pixel are extracted in the first two layers, and then the reconstructed pixels are obtained by the last layer.

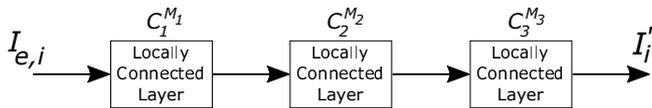


Fig. 4: Proposed DNN-based plaintext-only attack

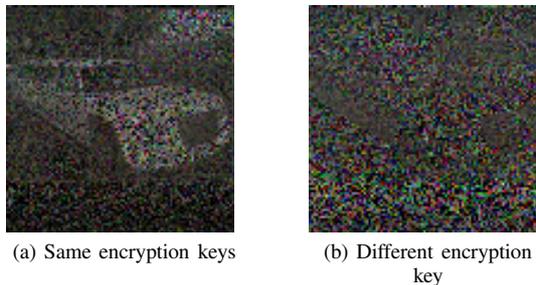


Fig. 5: Examples of reconstructed images from the images encrypted by the negative-positive transformation and color shuffling.

IV. EXPERIMENTS

A. Experimental Set-up

We employed STL-10 dataset, which contains 96×96 pixel color images and consists of 5,000 training images and 8,000 testing images [27].

In the experiment, the numbers of feature maps, M_1 , M_2 , and M_3 , were set to 8, 32, and 3, respectively.

The network was trained by using stochastic gradient descent (SGD) with momentum for 70 epochs, and used mean squared error (MSE), which compares the differences between the reconstructed images and the original ones, as a loss function. The learning rate was initially set to 0.1 and decreased by a factor of 10 at 40 and 60 epochs. We used a weight decay of 0.0005, a momentum of 0.9, and a batch size of 128.

B. Results and Discussions

Examples of reconstructed images under the use of same and different encryption keys are shown in Fig.5, where Fig.3(a) is the original one. The visual information of the reconstructed images was recovered by the proposed scheme if the images are encrypted under same encryption key, as shown in Fig. 5(a). This is because each image was encrypted with only one pattern, so the proposed attack can recognize the pattern and recover the visual information by comparing the difference between reconstructed images and original images.

TABLE II: Average SSIM of the reconstructed images compared with the original ones.

Key Conditions	Encryption	SSIM
Same encryption key	Step 2	0.1732
	Step 2 and 3	0.1715
Different encryption keys	Step 2	0.0424
	Step 2 and 3	0.0425

In comparison, the pixel-based encryption method has robustness against COA if the training images are encrypted by using different encryption keys. Therefore, the visual information cannot be recovered, as shown in Fig. 5(b).

Table II shows that the structural similarity (SSIM) values of the encrypted images under the use of same encryption key were much higher than under the use of different encryption keys.

V. CONCLUSION

This paper aimed to evaluate the safety of the pixel-based image encryption method in terms of robustness against COA. In addition, we proposed a novel DNN-based COA that aims to reconstruct the visual information of encrypted images. The effectiveness of the proposed attack was evaluated under two encryption key conditions: same encryption key, and different encryption keys. The experimental results showed that the proposed attack can recover the visual information of the encrypted images if images are encrypted under same encryption key. In contrast, it was proved that the pixel-based image encryption method has robustness against COA if images are encrypted under different encryption keys.

REFERENCES

- [1] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 1, Beijing, China, 2014, pp. 647–655.
- [2] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1, 2012, pp. 1097–1105.
- [3] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, April 2015, pp. 1–5.
- [4] A. Saxe, Y. Bansal, J.D., M. Advani, A. Kolchinsky, B. Tracey, and D. Cox, "On the information bottleneck theory of deep learning," in *International Conference on Learning Representations*, 2018.
- [5] T. Maekawa, A. Kawamura, Y. Kinoshita, and H. Kiya, "Privacy-preserving svm computing in the encrypted domain," in *Proceedings of APSIPA Annual Summit and Conference*, 2018, pp. 897–902.
- [6] M. Tanaka, "Learnable image encryption," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, May 2018, pp. 1–2.
- [7] I. Ito and H. Kiya, "One-time key based phase scrambling for phase-only correlation between visually protected images," *EURASIP Journal on Information Security*, vol. 2009, no. 841045, pp. 1–11, 2010.
- [8] Z. Tang, X. Zhang, and W. Lan, "Efficient image encryption with block shuffling and chaotic map," *Multimedia Tools Applications*, vol. 74, no. 15, pp. 5429–5448, 2015.
- [9] K. Kurihara, S. Shiota, and H. Kiya, "An encryption-then-compression system for jpeg standard," in *Picture Coding Symposium (PCS)*, 2015, pp. 119–123.
- [10] K. Kurihara, M. Kikuchi, S. Imaizumi, S. Shiota, and H. Kiya, "An encryption-then-compression system for jpeg/motion jpeg standard," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 98, no. 11, pp. 2238–2245, 2015.
- [11] T. Chuman, K. Kurihara, and H. Kiya, "On the security of block scrambling-based etc systems against jigsaw puzzle solver attacks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2157–2161.
- [12] —, "On the security of block scrambling-based etc systems against extended jigsaw puzzle solver attacks," *IEICE Transactions on Information and Systems*, vol. E101-D, no. 1, 2017.

- [13] T. Chuman, W. Sirichotedumrong, and H. Kiya, "Encryption-then-compression systems using grayscale-based image encryption for jpeg images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1515–1525, 2019.
- [14] W. Sirichotedumrong and H. Kiya, "Grayscale-based block scrambling image encryption using ycbcr color space for encryption-then-compression systems," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e7, 2019.
- [15] V. Itier, P. Puteaux, and W. Puech, "Recompression of jpeg crypto-compressed images without a key," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. [Online]. Available: <http://dx.doi.org/10.1109/TCSVT.2019.2894520>
- [16] M. T. Gaata and F. F. Hantoosh, "An efficient image encryption technique using chaotic logistic map and rc4 stream cipher," *International Journal of Modern Trends in Engineering and Research*, vol. 3, no. 9, pp. 213–218, 2016.
- [17] T. Araki, J. Furukawa, Y. Lindell, A. Nof, and K. Ohara, "High-throughput semi-honest secure three-party computation with an honest majority," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 805–817. [Online]. Available: <http://doi.acm.org/10.1145/2976749.2978331>
- [18] T. Araki, A. Barak, J. Furukawa, T. Lichter, Y. Lindell, A. Nof, K. Ohara, A. Watzman, and O. Weinstein, "Optimized honest-majority mpc for malicious adversaries – breaking the 1 billion-gate per second barrier," in *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017, pp. 843–862.
- [19] W. Lu, S. Kawasaki, and J. Sakuma, "Using fully homomorphic encryption for statistical analysis of categorical, ordinal and numerical data," *IACR Cryptology ePrint Archive*, vol. 2016, p. 1163, 2016.
- [20] Y. Aono, T. Hayashi, L. Phong, and L. Wang, "Privacy-preserving logistic regression with distributed data sources via homomorphic encryption," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 8, pp. 2079–2089, 2016.
- [21] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1310–1321. [Online]. Available: <http://doi.acm.org/10.1145/2810103.2813687>
- [22] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333–1345, May 2018.
- [23] L. Phong and T. Phuong, "Privacy-preserving deep learning for any activation function," *CoRR*, vol. abs/1809.03272, 2018. [Online]. Available: <http://arxiv.org/abs/1809.03272>
- [24] N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," Tech. Rep., February 2016.
- [25] Y. Wang, J. Lin, and Z. Wang, "An efficient convolution core architecture for privacy-preserving deep learning," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.
- [26] W. Sirichotedumrong, T. Maekawa, Y. Kinoshita, and H. Kiya, "Privacy-preserving deep neural networks with pixel-based image encryption considering data augmentation in the encrypted domain," in *IEEE International Conference on Image Processing (ICIP)*, September 2019, to be presented. [Online]. Available: <http://arxiv.org/abs/1905.01827>
- [27] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, vol. 15, 2011, pp. 215–223.