Image Transformation Network for Privacy-Preserving Deep Neural Networks and Its Security Evaluation

1st Hiroki Ito
Department of Computer Science
Tokyo Metropolitan University
Tokyo, Japan
ito-hiroki2@ed.tmu.ac.jp

2nd Yuma Kinoshita
Department of Computer Science
Tokyo Metropolitan University
Tokyo, Japan
ykinoshita@tmu.ac.jp

3rd Hitoshi Kiya
Department of Computer Science
Tokyo Metropolitan University
Tokyo, Japan
kiya@tmu.ac.jp

***Abstract*—We propose a transformation network for generating visually-protected images for privacy-preserving DNNs. The proposed transformation network is trained by using a plain image dataset so that plain images are transformed into visually protected ones. Conventional perceptual encryption methods have a weak visual-protection performance and some accuracy degradation in image classification. In contrast, the proposed network enables us not only to strongly protect visual information but also to maintain the image classification accuracy that using plain images achieves. In an image classification experiment, the proposed network is demonstrated to strongly protect visual information on plain images without any performance degradation under the use of CIFAR datasets. In addition, it is shown that the visually protected images are robust against a DNN-based attack, called inverse transformation network attack (ITN-Attack) in an experiment.**

***Index Terms*—deep neural network, privacy preserving, visual protection**

I. Introduction

The spread use of deep neural networks (DNNs) has greatly contributed to solving complex tasks for many applications [1], [2], including privacy-sensitive/security-critical ones such as facial recognition and medical image analysis. Recently, it has been very popular for data owners to utilize cloud servers to compute and process a large amount of data instead of using local servers. However, there are risks of data leakage in the cloud environment [3]. Because application users (i.e. clients) want to avoid the risks, privacy-preserving DNNs have become an urgent challenge. In this paper, we focus on protecting visual information on images before uploading them to cloud environments.

Perceptual encryption generates images that can be directly applied to various image processing algorithms, but information theory-based encryption (like RSA and AES) generates a ciphertext. In the past years, various perceptual encryption methods have already been proposed [4]–[19]. In these methods, there are only three methods for privacy-preserving DNNs: Tanaka's method [15], a pixel-based encryption method [16], [17], and a generative adversarial network (GAN)-based method using an image transformation network [19]. However, the use of the methods degrades the performance of DNNs, compared with the use of plain images.

For such reasons, in this paper, we propose a transformation network for generating visually-protected images for privacy-preserving DNNs. The proposed network transforms a plain image into a visually-protected one. The proposed network is trained so that generated images reduce the loss value of a classification model.

Experiments using the CIFAR-10 and 100 datasets [20] show that the proposed network enables us not only to protect visual information on plain images but also to maintain the performance of DNNs. Furthermore, we demonstrate that visual information on plain images can not be restored from visually-protected images by a DNN-based attack, called inverse transformation network attack (ITN-Attack).

II. Proposed Transformation Network

A. Overview

Figure 1 illustrates the framework that we assume in this paper. In this framework, transformation network h_θ is public to clients, and classification model ψ is available on a cloud server. The client sends visually-protected images generated by using h_θ to the cloud server. The cloud server classifies the images by using model ψ and returns the results to the client. In this framework, the cloud server has no visual information on plain images, so visual information is protected even if the cloud server is not trusted.

B. Training Transformation Network

The training procedure of the proposed transformation network is illustrated in Fig. 2, where $X = \{x_1, x_2, \ldots, x_m\}$ is an input plain image set, $\hat{X} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m\}$ is an output image set from the transformation network, i.e., $\hat{x}_i = h_\theta(x_i)$, $Y = \{y_1, y_2, \ldots, y_m\}$ is a one-hot encoded target label set, and $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m\}$ is an output label set from a classification network, i.e., $\hat{y}_i = \psi(\hat{x}_i)$. One-hot encoded

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

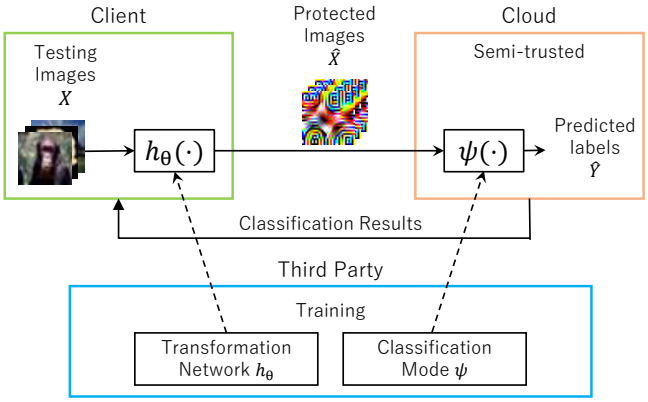


Fig. 1: Framework of proposed scheme

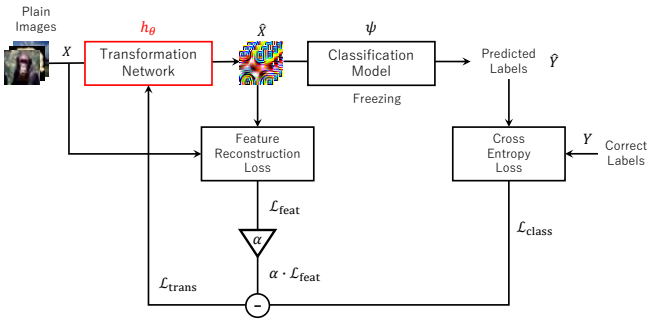


Fig. 2: Training process of transformation network h_θ

label, $y_i = (y_i(1), y_i(2), \ldots, y_i(c))$ and output label $\hat{y}_i = (\hat{y}_i(1), \hat{y}_i(2), \ldots, \hat{y}_i(c))$ meet

$$y_i(j) \in \{0, 1\}, \text{and} \sum_{j=1}^{c} y_i(j) = 1. \tag{1}$$

and

$$0 \leq \hat{y}_i(j) \leq 1, \text{and} \sum_{j=1}^{c} \hat{y}_i(j) = 1, \tag{2}$$

respectively, where c is the number of classes. The proposed network converts images to visually protected ones. Network h_θ is trained so that generated images reduce the loss value of model ψ.

To train network h_θ with parameter θ by using a plain input image x_i and its one-hot encoded target label y_i, loss function $\mathcal{L}_{\text{trans}}$ is minimized as

$$\underset{\theta}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_{\text{trans}}(x_i, h_\theta(x_i), y_i), \tag{3}$$

with

$$\mathcal{L}_{\text{trans}}(x_i, \hat{x}_i, y_i) = \mathcal{L}_{\text{class}}(\hat{x}_i, y_i) - \alpha \cdot \mathcal{L}_{\text{feat}}(x_i, \hat{x}_i), \tag{4}$$

where $\mathcal{L}_{\text{class}}$ denotes a classification loss function, which is used to classify visually protected images correctly, $\mathcal{L}_{\text{feat}}$ is a feature reconstruction loss function to be used for visually protecting input images, and $\alpha \in \mathbb{R}$ is a weight of $\mathcal{L}_{\text{feat}}$.

In this paper, $\mathcal{L}_{\text{class}}$ is given by the cross-entropy loss. Therefore, $\mathcal{L}_{\text{class}}$ is calculated by using $\hat{y}_i(j)$ as

$$\mathcal{L}_{\text{class}}(\hat{x}_i, y_i) = -\sum_{j=1}^{c} y_i(j) \log \hat{y}_i(j). \tag{5}$$

$\mathcal{L}_{\text{feat}}$ is also given by

$$\mathcal{L}_{\text{feat}}(x_i, \hat{x}_i) = \frac{1}{C_k H_k W_k} \|\phi_k(\hat{x}_i) - \phi_k(x_i)\|_2^2, \tag{6}$$

where $\phi_k(x)$ is a feature map with a size of $C_k \times H_k \times W_k$ obtained by the k-th layer of a network when image x is fed [21].

C. Robustness against DNN-based Attacks

One of the state-of-the-art attacks is a GAN (generative adversarial network)-based one [22]. The GAN-based attack may enable us to estimate visual information on plain images from visually-protected images without a correct pair set of plain images and protected images in general. Although, in our scheme (see Fig. 1), attackers can easily prepare a correct set because h_θ is open to the public. Therefore, attackers can create an inverse transformation network more efficiently by using a correct pair set for estimating visual information on plain images. In this paper, we tried to train an inverse transformation network by using a correct pair set. We call it an inverse transformation network attack (ITN-Attack). Even when ITN-Attack is applied to protected images generated by using h_θ, the protected ones will be shown to be robust enough against ITN-Attack in an experiment.

III. Simulations

We evaluated the proposed transformation network in terms of classification accuracy and visual protection performance. Robustness against ITN-Attack was also evaluated.

A. Evaluating Transformation Network Performance

We used U-Net [23] and ResNet-20 [24] as transformation network h_θ and classification model ψ, respectively. Also, we used the CIFAR-10 and 100 datasets [20]. Each dataset consists of a training set with 50,000 images and a test set with 10,000. In experiments using CIFAR-10, we utilized 45,000 images in the training set to train both ψ and h_θ, and the other 5,000 images were used as validation data. In contrast, we utilized 47,500 images for training both networks in experiments using CIFAR-100, and the other 2,500 images were used as validation data. The test set of each dataset was also utilized for evaluating the performance of the proposed method. In addition, standard data-augmentation methods, i.e., random crop and horizontal flip, were performed in the training.

All networks were trained for 200 epochs, by using the stochastic gradient descent (SGD) with a weight decay of 0.0005 and a momentum of 0.9. The learning rate was initially set to 0.1 and it was multiplied by 0.2 at 60, 120, and 160 epochs. The batch size was 128. After the training, we selected

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

TABLE I: Classification accuracy (%)

Method		CIFAR-10	CIFAR-100
Proposed	$\alpha = 0.005$	**91.72**	**70.78**
	$\alpha = 0.01$	**91.41**	**70.08**
	$\alpha = 0.05$	89.63	42.91
	$\alpha = 0.1$	39.92	1.00
Plain image		91.23	67.9
Tanaka [15]		85.18	60.08
Pixel-based [16], [17]		90.99	60.50

the network that provided the lowest loss value under the use of the validation set.

Figure 3 shows an example of visually protected images generated from ten test images in CIFAR-100, by using h_θ, where the top row shows plain images and the second top row to bottom row shows images generated with the parameters $\alpha = 0$, 0.005, and 0.01 in Eq. (4).

From the figure, the generated images had almost no visual information on the plain images when $\alpha \geq 0.005$. Also, in the case of $\alpha = 0$, the generated images were not visually protected, since a loss for visually protecting input images ($\mathcal{L}_{\rm feat}$) did not work. Also, all protected images have a similar pattern. Thus, the protected images have almost no visual information on the plain images in addition to the high classification accuracy.

Table I shows the classification accuracy when the generated images were protected. From the table, when $\alpha \leq 0.01$, the proposed network provided higher classification accuracy than conventional methods. The reason that the accuracy improved is that the proposed network increases the total number of parameters due to the use of the transformation network.

B. Evaluating Robustness against ITN-Attack

In Fig. 4, images estimated by using the inverse transformation model are illustrated together with the corresponding plain images and the visually protected ones. The inverse transformation model was trained by using h_θ trained with $\alpha = 0.005$. To evaluate the error of the estimation, peak signal-to-noise ratio (PSNR) between estimated images and the plain images were also calculated (see the bottom of each image). From Fig. 4, the estimated images had almost no visual information on the plain images and most estimated images had low PSNR values.

Figure 5 illustrates PSNR values calculated by using the 10,000 images in the test set of the CIFAR-100 dataset. The figure shows that the estimated images still had low PSNR values. In addition, all of the 10,000 estimated images were confirmed to have no visual information on plain images as well as in Fig. 4. From these results, visually protected images are robust against ITN-Attack.

IV. Conclusion

In this paper, we proposed a transformation network for generating visually-protected images for privacy-preserving DNNs. The proposed network enables us not only to protect visual information on plain images but also to maintain high classification accuracy. Experimental results demonstrated that images generated by the proposed transformation network have almost no visual information. We also confirmed that the visually-protected images are robust enough against ITN-Attack.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[2] 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 - Volume 32*, 2014, pp. 647–655.

[3] C.-T. Huang, L. Huang, Z. Qin, H. Yuan, L. Zhou, V. Varadharajan, and C.-C. J. Kuo, "Survey on securing data storage in the cloud," *APSIPA Transactions on Signal and Information Processing*, vol. 3, pp. 1–17, 2014.

[4] I. Ito and H. Kiya, "One-time key based phase scrambling for phase-only correlation between visually protected images," *EURASIP Journal on Information Security*, pp. 841 045–841 056, Dec. 2009.

[5] B. Ferreira, J. Rodrigues, J. Leitao, and H. Domingos, "Privacy-preserving content-based image retrieval in the cloud," in *2015 IEEE 34th Symposium on Reliable Distributed Systems (SRDS)*, Sept. 2015, pp. 11–20.

[6] J. Zhou, X. Liu, O. Au, and Y. Tang, "Designing an efficient image encryption-then-compression system via prediction error clustering and random permutation," *Information Forensics and Security, IEEE Transactions on*, vol. 9, pp. 39–50, Jan. 2014.

[7] Y. Zhang, B. Xu, and N. Zhou, "A novel image compression-encryption hybrid algorithm based on the analysis sparse representation," *Optics Communications*, vol. 392, no. C, pp. 223–233, 2017.

[8] K. Kurihara, S. Imaizumi, S. Shiota, and H. Kiya, "An encryption-then-compression system for lossless image compression standards," *IEICE Transactions on Information and Systems*, vol. E100.D, no. 1, pp. 52–56, 2017.

[9] 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, Jan. 2019.

[10] 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, June 2019.

[11] T. Maekawa, A. Kawamura, T. Nakachi, and H. Kiya, "Privacy-preserving support vector machine computing using random unitary transformation," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E102.A, no. 12, pp. 1849–1855, 2019.

[12] A. Kawamura, Y. Kinoshita, T. Nakachi, S. Shiota, and H. Kiya, "A Privacy-Preserving Machine Learning Scheme Using EtC Images," *arXiv e-prints*, p. arXiv:2007.08775, Jul. 2020.

[13] 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*, vol. 30, no. 3, pp. 646–660, 2020.

[14] S. Beugnon, P. Puteaux, and W. Puech, "Privacy protection for social media based on a hierarchical secret image sharing scheme," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sept. 2019, pp. 679–683.

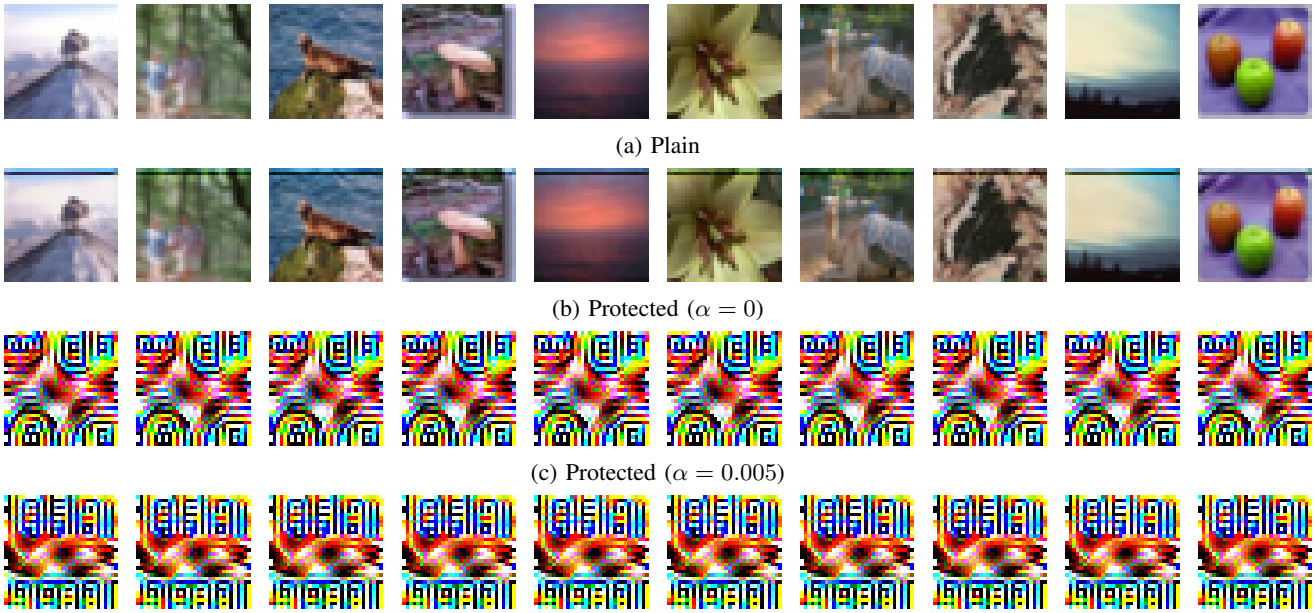
[15] M. Tanaka, "Learnable image encryption," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, May 2018, pp. 1–2.

[16] 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 *2019 IEEE International Conference on Image Processing (ICIP)*, Sept. 2019, pp. 674–678.

[17] W. Sirichotedumrong, Y. Kinoshita, and H. Kiya, "Pixel-based image encryption without key management for privacy-preserving deep neural networks," *IEEE Access*, vol. 7, pp. 177 844–177 855, 2019.

[18] 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, pp. 213–218, 2016.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE



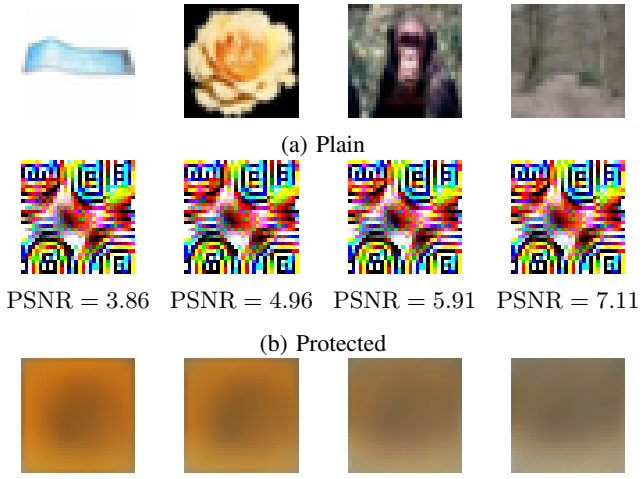
(a) Plain

(b) Protected ($\alpha = 0$)

(c) Protected ($\alpha = 0.005$)

(d) Protected ($\alpha = 0.01$)

Fig. 3: Visually protected images generated by proposed transformation network trained with ResNet-20



(a) Plain

(b) Protected

(c) Estimated

Fig. 4: Estimated images by inverse transformation network with h_θ trained with CIFAR-100 and $\alpha = 0.005$

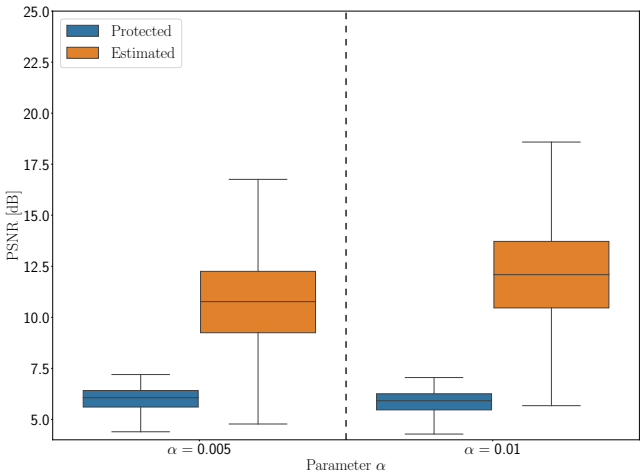


Fig. 5: PSNR values of estimated images. Boxes span from first to third quartile, referred to as Q_1 and Q_3, and whiskers show maximum and minimum values in range of $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$. Band inside box indicates median. Outliers are not indicated.

[19] W. Sirichotedumrong and H. Kiya, "A GAN-Based Image Transformation Scheme for Privacy-Preserving Deep Neural Networks," *arXiv e-prints*, p. arXiv:2006.01342, Jun. 2020.

[20] A. Krizhevsky, "Learning multiple layers of features from tiny images," Technical Report, 2009. [Online]. Available: https://www.cs.toronto.edu/ kriz/cifar.html

[21] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *Lecture Notes in Computer Science*, pp. 694–711, 2016.

[22] K. Madono, M. Tanaka, M. Onishi, and T. Ogawa, "An adversarial attack to learnable encrypted images (in japanese)," in *22nd IEICE Symposium on Image Recognition and Understanding*, 2019.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE