

Block-wise Scrambled Image Recognition Using Adaptation Network

Koki Madono^{1,2}, Masayuki Tanaka², Masaki Onishi², Tetsuji Ogawa^{1,2}

¹Department of Communications and Computer Engineering, Waseda University

²The National Institute of Advanced Industrial Science and Technology

Abstract

In this study, a perceptually hidden object-recognition method is investigated to generate secure images recognizable by humans but not machines. Hence, both the perceptual information hiding and the corresponding object recognition methods should be developed. Block-wise image scrambling is introduced to hide perceptual information from a third party. In addition, an adaptation network is proposed to recognize those scrambled images. Experimental comparisons conducted using CIFAR datasets demonstrated that the proposed adaptation network performed well in incorporating simple perceptual information hiding into DNN-based image classification.

Introduction

Cloud-based image analysis services, such as Google Cloud (Alphabet Inc.) and Microsoft Azure (Microsoft Corp.), have become extremely powerful and easy to use. These systems, however, can be further improved in terms of privacy issues. For example, when a client transfers an image to a cloud server, a third party can view the image. It is noteworthy that encrypted communications or secure communications are insufficient because the image data should be decrypted to analyze at the cloud. Deep neural networks (DNNs) have been widely used in image processing (LeCun, Bengio, and Hinton 2015; Lowe 1999; He et al. 2015); they are important in cloud-based image analysis. In general, DNNs require a large number of images for training. However, it becomes a critical problem when the cloud-based image analysis is conducted based on privacy-sensitive images, because the client has to provide such images to a third party for machine learning. Homomorphic encryption (Lagendijk, Erkin, and Barni 2013; Lu, Kawasaki, and Sakuma 2016) may address such a problem. However, it is not feasible from the viewpoints of computational and memory costs. Furthermore, mathematical operations in homomorphic encryptions are limited.

Table 1 summarizes the accessibility of humans and machines to perceptual information. Both humans and machines can access plain images. Although these images are

Table 1: Accessibility of human and machine to perceptual information. Difficulty of attacks is also listed as well.

	plain image	scrambled image	homographic encryption
human	✓		
machine	✓	✓	
attacker	easy	medium	difficult

easy to use, attackers can easily access them. Meanwhile, homographic encryptions can hide perceptual information; however, in general, it is difficult to use as aforementioned. Therefore, in the present study, a framework for hiding perceptual information from humans while machines continue accessing the information is developed.

Hence, the present study focuses on block-wise image scrambling as a simple perceptual information hiding method. Several studies regarding block-wise image scrambling have been conducted, such as learnable encryption (LE) (Tanaka 2018) and encryption-then-compression (EtC) (Chuman, Sirichotedumrong, and Kiya 2018). It is noteworthy that the scrambled images are not exactly encrypted although the perceptual information can be hidden. In addition, existing DNN architectures, which are proven effective in image recognition (He et al. 2015; Simonyan and Zisserman 2014; Yamada et al. 2018), do not assume the scrambled images as inputs, indicating that another component is required to recognize such images.

Hence, in this study, the LE algorithm is extended to increase the security level of perceptually hidden images and introduce the adaptation network for recognizing the scrambled images. Experimental verification conducted using CIFAR datasets (Lowe 1999) demonstrates that images with block-wise scrambling can be recognized by integrating the proposed adaptation network with the existing DNN-based classifier (He et al. 2015; Simonyan and Zisserman 2014; Yamada et al. 2018). Insights from the present study can contribute to the generation of images recognizable by humans but not machines.

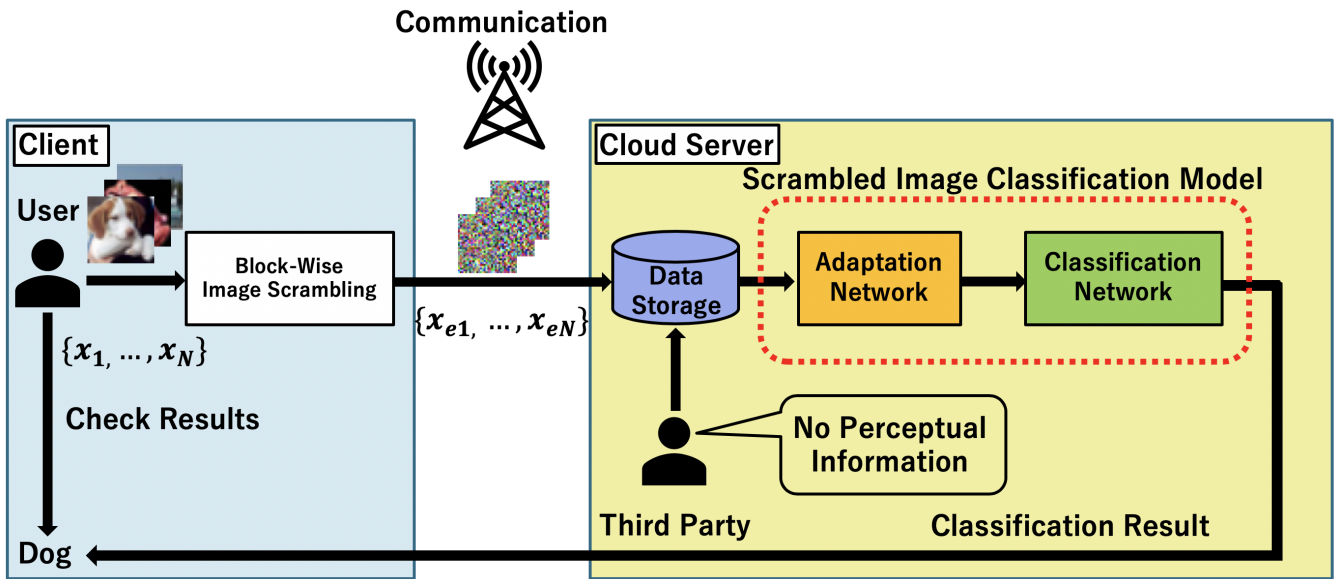


Figure 1: Conceptual image of cloud-based image analysis with image scrambling. Client can access the original image and object class in it. Third party for image analysis can access object class and perceptually hidden images for training image classifiers on cloud.

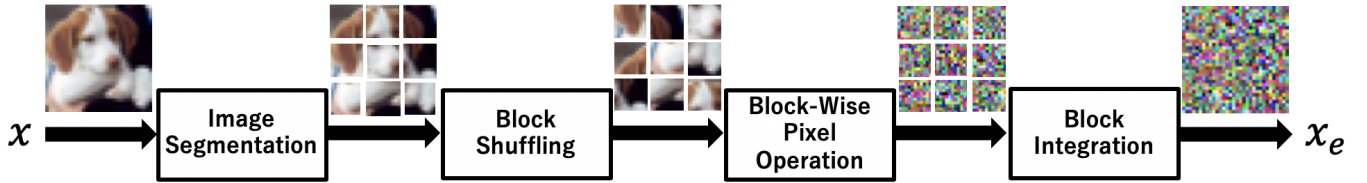


Figure 2: Processing pipeline of block-wise image scrambling. Image is segmented into blocks with $B \times B$ pixels. Block locations are shuffled, pixels in block are shuffled independently in every block, and shuffled blocks are concatenated.

Perceptual Information Hiding

Perceptual information hiding, in which perceptual information is hidden from humans but can be accessed by machines, is introduced to DNN-based image recognition. Figure 1 illustrates an assumed cloud-based image recognition, in which perceptually hidden images are obtained by block-wise image scrambling and used as inputs to a classifier on a cloud. During the classifier training, a client uploads scrambled images together with annotations. A third party of model developer then trains the model using the uploaded images. Because naive DNN-based image classifiers do not use scrambled images as inputs, the adaptation network is introduced to manage such concealed data.

Once the scrambled image classification models are trained, the assumed service can be used while hiding the perceptual information from a third party. The third party, however, can analyze the adaptation network to understand the scrambling process. Therefore, the use of adaptation networks is not perfect in perceptual information hiding but still offers some advantages because effective algorithms for reconstructing original images from block-wise scrambled images are still available, to the best of our knowledge.

Block-Wise Image Scrambling

This section describes the processing pipeline of block-wise image scrambling (Tanaka 2018; Chuman, Sirichotedumrong, and Kiya 2018). Figure 2 illustrates the pipeline. In the developed system, an input image is first divided into blocks with $B \times B$ pixels, where the number of yielded blocks is N . For example, the segmentation of a 32×32 pixel image into

blocks with 4×4 pixels yields 64 blocks. The positions of the segmented blocks are then shuffled (i.e., block shuffling). In addition, pixels in each block are shuffled with security keys, where different keys are applied to every blocks (i.e., block-wise pixel shuffling). Finally, the blocks with the shuffled pixels are concatenated to obtain the resulting scrambled image.

Table 2 summarizes block-wise image scrambling algorithms. The developed block-wise scrambling can be regarded as an extension of the existing LE algorithm (Tanaka 2018). Therefore, the developed algorithm is referred to as the extended learnable encryption (ELE).

Security Evaluation


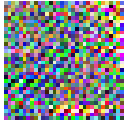

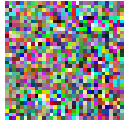
This subsection discusses the security levels of the block-wise image scrambling algorithms using the key space, which refers to the set of all possible permutations of a key.

The LE shuffles pixels and negative-positive transforms with the same key in every block. Each block is split into four upper bits and four lower bits to generate three to six channels of blocks. Subsequently, the pixels are randomly shuffled, yielding N_{ps} combinations. The intensity of randomly selected pixels are then reversed, yielding N_{np} combinations. The key space of the LE is calculated as

$$\begin{aligned}
 O(\text{LE}) &= N_{ps} \cdot N_{np} \\
 &= (4^2 \cdot 6)! \cdot 2^{4^2 \cdot 6} \\
 &= 96! \cdot 2^{96}.
 \end{aligned} \tag{1}$$

The EtC comprises block rotation followed by inversion ($N_{r\&i}$), negative-positive transform (N_{np}), color component

Table 2: Summary of block-wise scrambling algorithms. Here, each block has $B \times B$ pixels and the number of blocks is N .

	plain image	LE (Tanaka 2018)	EtC (Chuman <i>et al.</i> 2018)	ELE (proposed)
image example				
block key		common	different	different
block shuffling			✓	✓
block-wise pixel operation		pixel shuffling, negative-positive transform	block rotation & inversion, negative-positive transform, color component shuffling	pixel shuffling, negative-positive transform
key space	0	$(B^2 \cdot 6)! \cdot 2^{B^2 \cdot 6}$	$8^N \cdot 2^N \cdot 6^N \cdot N!$	$\{(B^2 \cdot 6)! \cdot 2^{B^2 \cdot 6}\}^N \cdot N!$

shuffling (N_{col}), and block location shuffling (N_{bs}) with the same key in every block. The key space of the EtC is calculated as

$$\begin{aligned}
 O(\text{EtC}) &= N_{\text{r\&i}} \cdot N_{\text{np}} \cdot N_{\text{col}} \cdot N_{\text{bs}} \\
 &= (4 \cdot 2)^{64} \cdot 2^{64} \cdot (3!)^{64} \cdot 64! \\
 &= 8^{64} \cdot 2^{64} \cdot 6^{64} \cdot 64!.
 \end{aligned} \tag{2}$$

The ELE comprises block-wise pixel shuffling with different keys in every block (N_{dps}) and block location shuffling (N_{bs}). The key space of the ELE is calculated as

$$\begin{aligned}
 O(\text{ELE}) &= N_{\text{dps}} \cdot N_{\text{bs}} \\
 &= \{(4^2 \cdot 6)! \cdot 2^{4^2 \cdot 6}\}^{64} \cdot 64! \\
 &= (96! \cdot 2^{96})^{64} \cdot 64!
 \end{aligned} \tag{3}$$

Eventually, the key spaces for the LE, EtC, and ELE are as follows:

$$O(\text{ELE}) \gg O(\text{EtC}) \gg O(\text{LE}). \tag{4}$$

This indicates that the ELE has a larger key space than the LE and EtC.

Adaptation Network for Block-Wise Scrambled Image Recognition

In this section, proposes the adaptation network to recognize the block-wise scrambled images, considering the processing pipeline of block-wise scrambling as shown in Fig. 2.

The proposed adaptation network comprises three parts: block-wise sub-networks, a learnable pseudo permutation matrix, and a pixel shuffling layer (Shi *et al.* 2016). The schematic diagram of the developed system and detailed information on network architecture for each component are illustrated in Figs. 3 and 4, respectively.

First, a scrambled image \mathbf{x}_e is segmented into N blocks with $B \times B$ pixels as $\{\mathbf{x}_{e_1}, \mathbf{x}_{e_2}, \dots, \mathbf{x}_{e_N}\}$, where \mathbf{x}_{e_b}

represents a block (i.e., segmented image). Each block is transformed by the corresponding block-wise sub-network $f(\mathbf{x}_{e_b}; \boldsymbol{\theta}_b)$, which is separately trained for each block. This process aims at managing ELE-based image scrambling i.e., block-wise scrambling with different keys. The feature maps of the block-wise sub-networks are then integrated and the pseudo permutation matrix is applied to the integrated feature map. This operation corresponds to inverse block shuffling. Subsequently, the shuffled pixels are aligned on the pixel shuffling layer and then used as an input to the classification network. In theory, if the permutation matrix for block shuffling is known, its inverse matrix can be used to solve block shuffling. Such a permutation matrix, however, is unknown. Instead of estimating the inverse permutation matrix, we introduce pseudo permutation and train the corresponding matrix U such that its elements can be sparse. Hence, the L_{1-2} penalty (Lyu *et al.* 2019; Esser, Lou, and Xin 2013; Yin *et al.* 2015) is added to the loss function for training. In addition, all elements of U should be non-negative and the sum of each column and that of each row should be one. Such constraints, however, are not considered herein.

The feature map of the plain image is expected to be spatially smooth. Additionally, the feature map of the adaptation network should exhibit the same characteristics as those of the natural image. Therefore, the penalty with respect to spatial smoothness is added to the loss function.

The resulting loss function for training the adaptation network is written as

$$L = L_{CE} + \lambda_U L_U + \lambda_s L_s, \tag{5}$$

where L_{CE} denotes the cross entropy loss for classification (Mannor, Peleg, and Rubinstein 2005), L_U denotes the L_{1-2} penalty for the pseudo permutation matrix, and L_s denotes the spatial smoothness penalty for the feature map of the adaptation network. Here, λ_U and λ_s were empirically determined as $\lambda_U = 0.001$ and $\lambda_s = 0.1$, respectively.

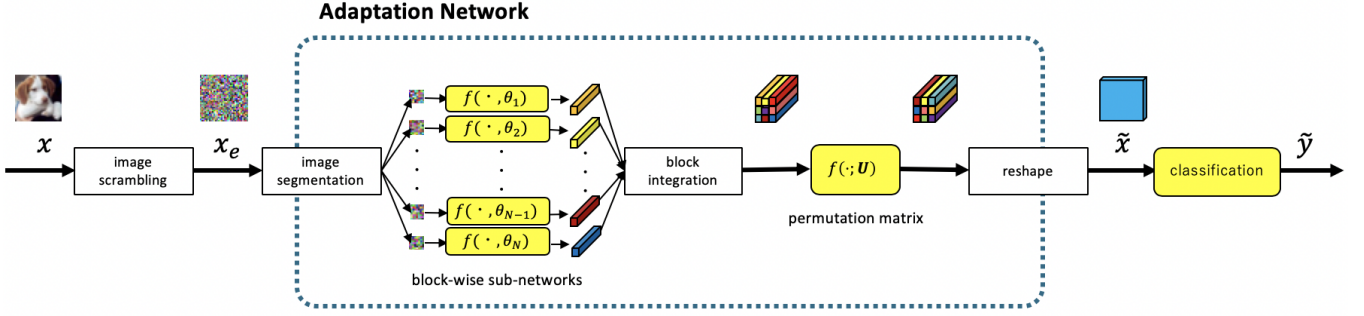


Figure 3: Network architecture of proposed adaptation network. Scrambled image is segmented into blocks with $B \times B$ pixels; feature maps of block-wise sub-networks, $\{f(x_{e_1}; \theta_1), \dots, f(x_{e_N}; \theta_N)\}$, are integrated; pseudo permutation matrix is applied to integrated feature map; and feature map is upscaled by pixel shuffle layer (Shi et al. 2016) and then used as input to classification network based on shakedrop. Colored boxes mean trainable processing.

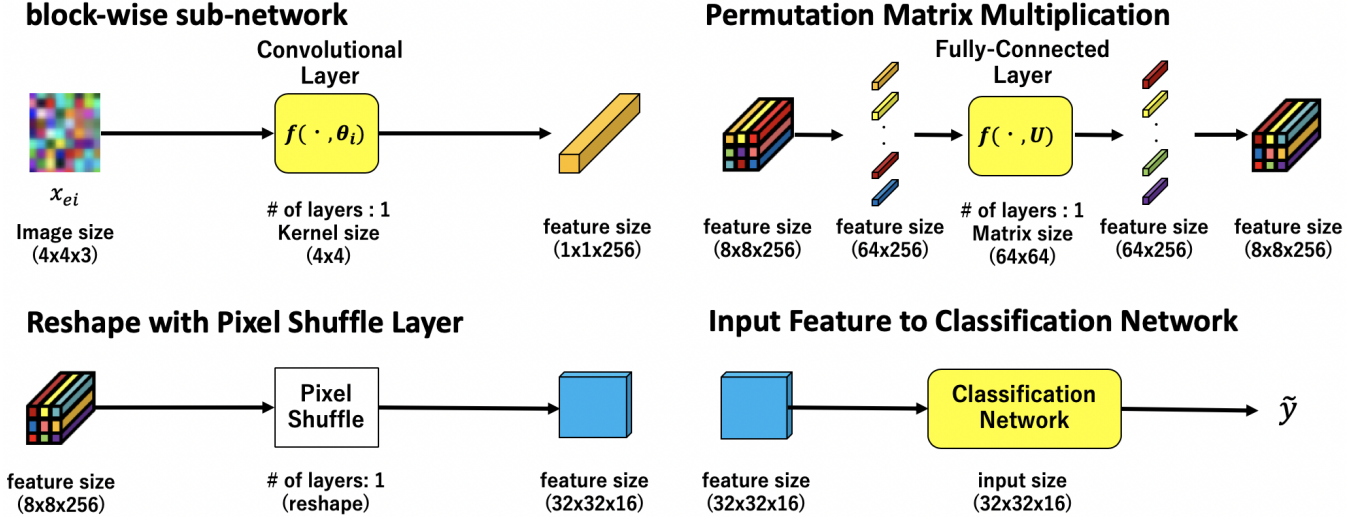


Figure 4: Detailed network architectures used. Block size and kernel size are the same in every block-wise sub-networks. Pseudo permutation matrix is applied to integrated feature map; Feature map is upscaled by pixel shuffle layer (Shi et al. 2016); and feature map is used as input to classification network (Yamada et al. 2018). Colored boxes express trainable processing.

The cross entropy loss L_{CE} is computed as

$$L_{CE} = -\frac{1}{M} \sum_{m=1}^M \sum_{k=1}^K t_{m,k} \log C_k(\tilde{x}_m; \theta_c), \quad (6)$$

where m and k denote the sample and class indexes, respectively; $t_{m,k}$ denotes the one-hot encoded label; $C_k(\tilde{x}_m; \theta_c)$ denotes the posterior probability for the k -th class of the classification network; \tilde{x}_m denotes the feature map of the adaptation network for the m -th sample; and θ_c denotes the parameters of the classification network.

The L_{1-2} penalty for the pseudo permutation matrix L_U is calculated as

$$L_U = \frac{1}{N \times N} \left\{ \sum_{i=1}^N \left[\sum_{j=1}^N |u_{i,j}| - \left(\sum_{j=1}^N u_{i,j}^2 \right)^{1/2} \right] + \sum_{j=1}^N \left[\sum_{i=1}^N |u_{i,j}| - \left(\sum_{i=1}^N u_{i,j}^2 \right)^{1/2} \right] \right\}, \quad (7)$$

where $u_{i,j}$ denotes the element of the matrix U whose size is $N \times N$.

The spatial smoothness penalty of the feature map L_s is calculated as

$$L_s = \frac{1}{M} \sum_{M=1}^M \left[L_{s,i}^{(H)} + L_{s,i}^{(V)} \right]$$

$$L_{s,i}^{(H)} = \frac{1}{H(W-1)C} \sum_{h=1}^H \sum_{w=1}^{W-1} \sum_{c=1}^C [\tilde{x}_i^{(H)}(h, w, c)]^2 \quad (8)$$


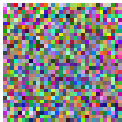

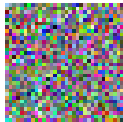
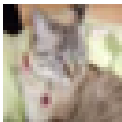
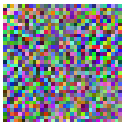

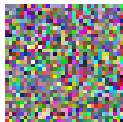

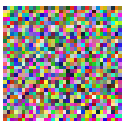

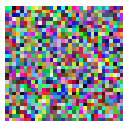
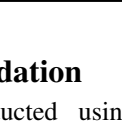
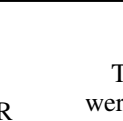
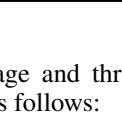
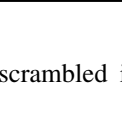
$$L_{s,i}^{(V)} = \frac{1}{(H-1)WC} \sum_{h=1}^{H-1} \sum_{w=1}^W \sum_{c=1}^C [\tilde{x}_i^{(V)}(h, w, c)]^2,$$

where H , W , and C denote the height, width, and number of channels in the feature map, respectively; $\tilde{x}_i^{(H)}(h, w, c)$ and $\tilde{x}_i^{(V)}(h, w, c)$ denote the horizontal and vertical forward difference of the i -th feature at (h, w, c) , respectively.

Table 3: Accuracy of scrambled image classification.

Dataset	Adaptation network	plain image	LE (Tanaka 2018)	EtC (Chuman <i>et al.</i> 2018)	ELE (proposed)
cifar-10	No AdaptNet	96.70%	94.94%	85.94%	67.10%
	LE-AdaptNet	95.64%	94.49%	80.16%	48.39%
	ELE-AdaptNet	85.32%	87.28%	89.09%	83.06%
cifar-100	No AdaptNet	83.59%	78.25%	61.90%	43.05%
	LE-AdaptNet	79.13%	75.48%	44.83%	7.19%
	ELE-AdaptNet	60.36%	71.30%	71.91%	62.97%

Table 4: Estimates and their posterior probabilities of scrambled image classification using different models.

Adaptation Network	plain image	LE (Tanaka 2018)	EtC (Chuman <i>et al.</i> 2018)	ELE (proposed)
Shakedown LE-AdaptNet + Shakedown ELE-AdaptNet + Shakedown				
	frog(1.00)	frog(1.00)	frog(1.00)	deer(0.65)
	frog(1.00)	frog(1.00)	frog(0.99)	deer(0.63)
Shakedown LE-AdaptNet + Shakedown ELE-AdaptNet + Shakedown				
	cat(1.00)	cat(1.00)	cat(0.99)	bird(0.28)
	cat(1.00)	cat(0.99)	cat(0.99)	bird(0.34)
Shakedown LE-AdaptNet + Shakedown ELE-AdaptNet + Shakedown				
	automobile(1.00)	automobile(1.00)	automobile(0.98)	automobile(0.44)
	automobile(1.00)	automobile(1.00)	automobile(0.99)	dog(0.55)
Shakedown LE-AdaptNet + Shakedown ELE-AdaptNet + Shakedown				
	automobile(0.97)	automobile(0.94)	automobile(1.00)	automobile(0.82)
	automobile(0.97)	automobile(0.94)	automobile(1.00)	automobile(0.82)

Experimental Validation

Experimental validation was conducted using CIFAR datasets to demonstrate the effectiveness of the proposed scrambled image recognition method¹. Three systems using the following adaptation networks were compared as follows:

- **No AdaptNet**: No adaptation network
- **LE-AdaptNet**: Adaptation network with block independent (shared) sub-networks (Tanaka 2018)
- **ELE-AdaptNet**: Adaptation network with block dependent sub-networks and block permutation matrix (proposed)

Because the existing **LE-AdaptNet** (Tanaka 2018) considers LE-based block-wise image scrambling, it shares the block-wise network among all blocks and does not require a pseudo permutation matrix.

¹<https://github.com/MADONOKOUKI/Block-wise-Scrambled-Image-Recognition>

The plain image and three types of scrambled images were evaluated as follows:

- **plain image** (i.e., no scrambling);
- learnable encryption (**LE**) (Tanaka 2018);
- encryption-then-compression (**EtC**) (Chuman, Sirichote-dumrong, and Kiya 2018); and
- extended learnable encryption (**ELE**) proposed in this study.

The shakedown network (Yamada et al. 2018) was exploited for the classification network with the same setting as that of the original implementation.

Experimental Setup

The datasets used for the evaluation were CIFAR-10 and CIFAR-100 (Lowe 1999). All images were converted into block-wise scrambled images and used as inputs to the adaptation network. The data augmentation was performed before block-wise scrambling. The mini-batch size was 128 during training and testing. The SGD with the Nesterov

was used as the optimizer, where the momentum was 0.9. The network was trained with 300 epochs of iterations. The learning rate was scheduled as 0.1 for 0-to-150 epochs, 0.01 for 150-to-225 epochs, and 0.001 for 225-to-300 epochs.

Experimental Results

Table 3 lists the accuracy of recognizing the plain images and three types of scrambled images, where the boldface represents the best accuracy. The proposed **ELE-AdaptNet** yielded the best accuracy for ELE- and EtC-based image scrambling. For LE-based image scrambling, the **LE-AdaptNet** (Tanaka 2018) performed better than the **ELE-AdaptNet** because the **LE-AdaptNet** is specialized to LE-based image scrambling. However, it is noteworthy that the LE-based image scrambling is worse in terms of security than the ELE-based image scrambling, as shown in Table 2. Either the LE or ELE can be selected as a perceptual information hiding method, considering the tradeoff between security level and recognition accuracy. If the security level is prioritized, the ELE-based approach should be selected.

Table 4 lists examples of the plain and scrambled images with the best matching object classes and corresponding posterior probabilities in parentheses. For the plain image and LE- and EtC-based scrambled images, the posterior probabilities are confident, regardless of the adaptation networks. For ELE-based image scrambling, the proposed **ELE-AdaptNet** yielded confident posterior probabilities, while the other networks did not perform as intended. This indicates that the proposed **ELE-AdaptNet** achieved reliable recognition, irrespective of image type. Furthermore, as the security level of the scrambled image increases, an adequate adaptation network is required for a reliable recognition.

Conclusion

The block-wise scrambling algorithm was introduced herein to hide perceptual information. In addition, an adaptation network was proposed to recognize such scrambled images. Experimental comparisons implied that the proposed block-wise scrambling and adaptation network could achieve simple perceptual information hiding on DNN-based image analysis.

Acknowledgements

This work was supported by JST CREST Grant Number JP-MJCR19F5.

References

Alphabet Inc. Google cloud. <https://cloud.google.com>.
Chuman, T.; Sirichotedumrong, W.; and Kiya, H. 2018. Encryption-then-compression systems using grayscale-based image encryption for jpeg images. *IEEE Transactions on Information Forensics and Security* 14:1515–1525.
Esser, E.; Lou, Y.; and Xin, J. 2013. A method for finding structured sparse solutions to nonnegative least squares problems with applications. *SIAM J. Imaging Sciences* 6:2010–2046.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778.
Legendijk, R. L.; Erkin, Z.; and Barni, M. 2013. Encrypted signal processing for privacy protection: Conveying the utility of homomorphic encryption and multiparty computation. *IEEE Signal Processing Magazine* 30:82–105.
LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521:436–444.
Lowe, D. G. 1999. Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision* 2:1150–1157.
Lu, W.; Kawasaki, S.; and Sakuma, J. 2016. Using fully homomorphic encryption for statistical analysis of categorical, ordinal and numerical data. *IACR Cryptology ePrint Archive* 2016:1163.
Lyu, J.; Zhang, S.; Qi, Y.; and Xin, J. 2019. Autosshuffenet: Learning permutation matrices via an exact lipschitz continuous penalty in deep convolutional neural networks. *ArXiv abs/1901.08624*.
Mannor, S.; Peleg, D.; and Rubinstein, R. Y. 2005. The cross entropy method for classification. In *ICML*.
Microsoft Corp. Microsoft azure. <https://azure.microsoft.com>.
Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1874–1883.
Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556*.
Tanaka, M. 2018. Learnable image encryption. *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)* 1–2.
Yamada, Y.; Iwamura, M.; Akiba, T.; and Kise, K. 2018. Shakedown regularization for deep residual learning. volume *abs/1802.02375*.
Yin, P.; Lou, Y.; He, Q.; and Xin, J. 2015. Minimization of ℓ_{1-2} for compressed sensing. *SIAM J. Scientific Computing* 37.