On the Transferability of Adversarial Examples between Encrypted Models

1st Miki Tanaka
Tokyo Metropolitan University
Tokyo, Japan
tanaka-miki@ed.tmu.ac.jp

2nd Isao Echizen
National Institute of Informatics (NII)
Tokyo, Japan
iechizen@nii.ac.jp

3rd Hitoshi Kiya
Tokyo Metropolitan University
Tokyo, Japan
kiya@tmu.ac.jp

***Abstract*—Deep neural networks (DNNs) are well known to be vulnerable to adversarial examples (AEs). In addition, AEs have adversarial transferability, namely, AEs generated for a source model fool other (target) models. In this paper, we investigate the transferability of models encrypted for adversarially robust defense for the first time. To objectively verify the property of transferability, the robustness of models is evaluated by using a benchmark attack method, called AutoAttack. In an image-classification experiment, the use of encrypted models is confirmed not only to be robust against AEs but to also reduce the influence of AEs in terms of the transferability of models.**

***Index Terms*—Adversarial example, Deep learning, Transferability**

I. Introduction

Deep neural networks (DNNs) have been deployed in many applications including security-critical ones such as biometric authentication and automated driving, but DNNs are vulnerable to adversarial examples (AEs), which are perturbed by noise to mislead DNNs without affecting human perception. In addition, AEs generated for a source model fool other (target) models, and this property is called adversarial transferability. This transferability allows attackers to use a substitute model to generate AEs that may also fool other target models, so reducing the influence of the transferability has become an urgent issue. Many studies have investigated both AEs and the transferability of AEs to build models robust against these attacks. In contrast, various methods for generating AEs have also been proposed to fool DNNs.

One of the methods for constructing models robust against AEs is to train models by using encrypted images [1]–[3], which was inspired by learnable image encryption [4]–[10]. The method was demonstrated to be robust against various AEs, but it has never been evaluated in terms of the transferability of AEs. Accordingly, in this paper, we aim to evaluate the method with encrypted images in terms of the transferability and encryption settings. In addition, the evaluation of encrypted models is carried out under the use of a benchmark attack method, referred to as AutoAttack, which was proposed to objectively evaluate the robustness of models against AEs. In an experiment, the use of encrypted models is verified not only to be robust against AEs but to also reduce the influence of the transferability between models.

II. Related Work

A. Adversarial examples

AEs are classified into three groups based on the knowledge of a particular model and training data available to the adversary: white-box, black-box, and gray-box. Under white-box settings [11]–[13], the adversary has direct access to the model, its parameters, training data, and defense mechanism. However, the adversary does not have any knowledge on the model, except the output of the model in black-box attacks [14]–[16]. Situated between white-box and black-box methods are gray-box attacks that imply that the adversary knows something about the system. With the development of AEs, numerous adversarial defenses have been proposed in the literature. Conventional defenses have been compared under a benchmark attack framework called AutoAttack [17].

Many studies [18]–[21] have investigated adversarial transferability. The transferability in these studies is classified into two groups, i.e., non-targeted and targeted transferability, in accordance with the objective of the adversarial attack. However, no adversarial defense method with encrypted models has ever been evaluated in terms of robustness against adversarial transferability.

B. AutoAttack

Many defenses against AEs have been proposed, but it is very difficult to judge the value of defense methods without an independent test. For this reason, AutoAttack [17], which is an ensemble of adversarial attacks used to test adversarial robustness objectively, was proposed as a benchmark attack. AutoAttack consists of four attack methods: Auto-PGD-cross entropy (APGD-ce) [17], APGD-target (APGD-t), FAB-target (FAB-t) [13], and Square Attack [15], as summarized in Table I. In this paper, we use these four attack methods to objectively evaluate the transferability of AEs.

III. Analysis of encrypted models

In this paper, the robustness of encrypted models is evaluated under various settings. Targets for comparison are summarized here.

TABLE I
ATTACK METHODS USED IN AUTOATTACK

Attack	Target (T)/ Non-target (N)	White-box (W)/ Black-box (B)
APGD-ce	N	W
APGD-t	T	W
FAB-t	T	W
Square	N	B

A. Type of model

Various models have been proposed for image classification tasks. The residual network (ResNet) [22] and very deep convolutional network (VGGNet) [23] use a convolutional neural network(CNN). In contrast, vision transformers (ViT) [24] do not. In previous work, the transferability between CNN models and ViT was mentioned to be lower than the transferability between CNN models [21]. In this paper, we use three CNN models, ResNet18, ResNet50, and VGG16, and we also use ViT to investigate the transferability of AEs between models in addition to encrypted CNN models.

B. Encrypted model

A block-wise transformation with secret keys was proposed for adversarial defense [1], where a model is trained by using encrypted images as below (see Fig. 1).

1) Each training image x is divided into blocks with a size of $M \times M$.
2) Every block in x is encrypted by a transformation algorithm with secret keys to generate encrypted image.
3) A model is trained by using the encrypted images to generate an encrypted model.
4) A query image is encrypted with key K, and an encrypted image is then input to the encrypted model to get an estimation result.

There are two parameters in steps 1) and 2), when encrypting a model: block size M and transformation algorithm. In [1], three transformation algorithms were proposed: pixel shuffling (SHF), bit flipping (NP), and format-preserving, Feistel-based encryption (FFX). Figure 2 shows an example of images encrypted from the original image with a size of 224 × 224 in Fig. 2(a) by using these three algorithms with $M = 16$.

In this paper, we evaluate the transferability of AEs between models including encrypted ones.

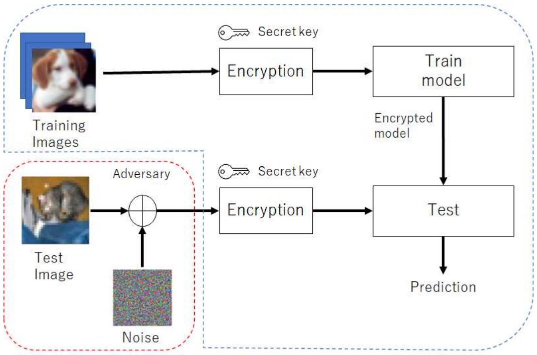

Fig. 1. Adversarial defense with encrypted model

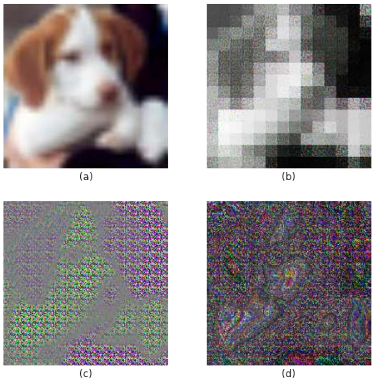

Fig. 2. Example of transformed images (M=16) (a): plain image, (b): SHF, (c): NP, (d): FFX

IV. EXPERIMENT

A. Experiment setup

In the experiment, we used four networks for image classification, ResNet18, ResNet50, VGG16, and ViT, to evaluate the transferability of AEs. In addition, ResNet18 was used for generating encrypted models where the above three transformation algorithms, SHF, NP, and FFX were applied to images in accordance with the steps in Sec Ⅲ B. The experiment was carried out on the CIFAR-10 dataset (with 10 classes), which consists of 60,000 color images (dimension of 3 × 32 × 32), where 50,000 of the images are for training and 10,000 for testing, and each class contains 6000 images. The images in the dataset were resized to 3 × 224 × 224 for fitting with pretrained ViT models. AEs were generated by using four attack methods under the l_∞ norm with $\epsilon = 8/255$ used in AutoAttack: APGD-ce, APGD-t, FAB-t, and Square. For white-box attacks, the adversary was assumed to be able to access the secret keys.

The transferability of AEs was evaluated by using two assessment criteria: accuracy (Acc) and attack success rate (ASR). Acc is given by

$$\mathrm{Acc} = \frac{100}{N} \sum_{k=1}^{N} \begin{cases} 1 & C_t(x_k^{adv}) = y_k \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where x_k^{adv} is an AE generated from an image x_k, y_k is a label of x_k, C_t is a target classifier, and N is the number of input images. Acc is in the range of [0, 100] and a higher value indicates that images are classified more correctly.

The ASR between a source classifier C_s and C_t is also given by

$$\mathrm{ASR} = \frac{100}{N_c} \sum_{k=1}^{N_c} \begin{cases} 1 & A_{C_t}(x_k, y_k) \wedge \{C_s(x_k) = y_k\} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

$$A_C(x, y) = \{C(x) = y\} \wedge \{C(x^{adv}) \neq y\}, \tag{3}$$

where N_c is the number of images correctly classified in both C_t and C_s. ASR is in the range of [0,100], and a lower value indicates that the transferability is lower.

B. Results

1) Transferability between CNN and ViT: As shown in Tables II and III, the transferability between plain models was evaluated in terms of ACC and ASR, respectively, where ResNet18 was chosen as the source model for which AEs were generated. From the tables, the AEs generated for the source model misled all CNN models including ResNet50 and VGG16. In contrast, they could not mislead the ViT models.

Tables IV and V show the results obtained when ViT was chosen as the source model. The AEs generated with the ViT models could not fool the CNN models successfully. Thus, the transferability between CNN and ViT models was confirmed to be low as described in [21].

TABLE II
TRANSFERABILITY BETWEEN CNN AND VIT MODELS (SOURCE: RESNET18, ACC)

Target	APGD-ce	APGD-t	FAB-t	Square
ResNet18	0.00	0.00	0.36	0
ResNet50	4.84	30.64	92.69	74.20
VGG16	43.39	62.81	91.59	85.91
ViT	68.92	93.95	98.99	95.26

TABLE III
TRANSFERABILITY BETWEEN CNN AND VIT MODELS (SOURCE: RESNET18, ASR)

Target	APGD-ce	APGD-t	FAB-t	Square
ResNet18	100.00	100.00	99.62	100.00
ResNet50	97.08	68.97	1.35	21.74
VGG16	54.39	33.06	0.69	7.68
ViT	32.16	5.49	0.04	4.09

TABLE IV
TRANSFERABILITY BETWEEN CNN AND VIT MODELS (SOURCE: VIT, ACC)

Target	APGD-ce	APGD-t	FAB-t	Square
ResNet18	85.45	89.34	93.60	87.04
ResNet50	86.28	89.18	93.45	83.89
VGG16	84.22	85.90	91.93	87.85
ViT	0.00	0.00	0.00	0.88

2) Transferability between CNN and encrypted CNN models: Next, the transferability between CNN and encrypted CNN models was confirmed experimentally as shown in Tables VI and VII, where ResNet18 was chosen as the source model for generating AEs as well, and ResNe18 was also encrypted for model protection. The use of encrypted models was verified to be effective for making the transferability of AEs weak. In particular, the use of FFX and a large block size enhanced the effect.

TABLE V
TTRANSFERABILITY BETWEEN CNN AND VIT MODELS (SOURCE: VIT, ASR)

Target	APGD-ce	APGD-t	FAB-t	Square
ResNet18	9.60	5.56	0.55	8.16
ResNet50	8.60	5.79	0.57	11.49
VGG16	9.51	7.97	0.35	5.87
ViT	100.00	100.00	100.00	99.11

TABLE VI
TRANSFERABILITY BETWEEN CNN AND ENCRYPTED MODELS (SOURCE: RESNET18, TARGET: ENCRYPTED RESNET18, ACC)

transform	block size	APGD-ce	APGD-t	FAB-t	Square
SHF	4	2.71	22.16	92.43	80.12
SHF	8	6.25	37.42	93.00	76.61
SHF	16	41.84	75.88	92.06	80.12
NP	4	3.01	21.81	93.22	75.60
NP	8	2.75	21.88	93.22	75.31
NP	16	13.12	56.58	93.45	80.42
FFX	4	18.93	68.58	91.91	84.40
FFX	8	35.73	77.25	92.05	85.06
FFX	16	71.22	86.02	92.29	85.69

3) Transferability between encrypted models: The transferability between models encrypted under different conditions was evaluated as shown in Tables VIII and IX, where the parameters used for encryption were block size, transformation algorithm, and secret key, and the source model was encrypted with a block size of 4 and SHF. As shown in the tables, when different parameters were used for each model encryption, the transferability between encrypted models was reduced. In particular, the use of FFX and a large block size can reduce the transferability more between models as well.

V. CONCLUSION

In this paper, we investigated the transferability of models including encrypted ones. To objectively verify the transferability, four attack methods used in AutoAttack, were used to generate AEs from a source model. In the experiment, the use of encrypted models was confirmed not only to be robust against AEs but to also reduce the influence of the transferability between models. In addition, the use of FFX and a large block for image encryption was effective for making the transferability of a model weak.

ACKNOWLEDGMENTS

The research was partially supported by JST CREST (Grant Number JPMJCR20D3) and ROIS NII Open Collaborative Research 2022-(22S1401).

REFERENCES

[1] M. AprilPyone and H. Kiya, “Block-wise image transformation with secret key for adversarially robust defense,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2709–2723, 2021.

[2] ——, “Encryption inspired adversarial defense for visual classification,” in *Proc. of IEEE International Conference on Image Processing*, 2020, pp. 1681–1685.

TABLE VII
TRANSFERABILITY BETWEEN CNN AND ENCRYPTED MODELS (SOURCE: RESNET18, TARGET: ENCRYPTED RESNET18, ASR)

transform	block size	APGD-ce	APGD-t	FAB-t	Square
SHF	4	99.29	78.16	1.41	15.17
SHF	8	95.76	61.75	1.03	19.21
SHF	16	56.04	18.57	0.62	14.16
NP	4	99.22	78.77	1.11	20.42
NP	8	99.38	78.64	1.13	20.87
NP	16	88.08	40.74	0.58	14.95
FFX	4	81.81	27.87	2.51	10.86
FFX	8	63.34	18.11	1.98	9.83
FFX	16	24.37	8.18	1.29	8.77

TABLE VIII
TRANSFERABILITY BETWEEN ENCRYPTED MODELS (SOURCE: RESNET18 ENCRYPTED WITH SHF AND M=4, TARGET: ENCRYPTED RESNET18, ACC)

transform	block size	APGD-ce	APGD-t	FAB-t	Square
no (plain)	-	85.35	88.14	93.95	83.58
SHF (same key as target)	4	0.00	0.00	0.18	0.00
SHF (different key)	4	69.99	79.60	93.26	79.34
SHF	8	81.7	85.26	93.83	78.89
SHF	16	87.08	88.77	92.57	81.43
NP	4	86.76	88.81	94.11	79.64
NP	8	83.73	86.62	94.19	78.20
NP	16	90.41	91.34	93.90	83.43
FFX	4	90.31	92.03	93.21	86.40
FFX	8	91.54	92.18	93.20	87.37
FFX	16	90.60	91.15	92.74	87.02

TABLE IX
TRANSFERABILITY BETWEEN ENCRYPTED MODELS (SOURCE: RESNET18 ENCRYPTED WITH SHF AND M=4, TARGET: ENCRYPTED RESNET18, ASR)

transform	block size	APGD-ce	APGD-t	FAB-t	Square
no (plain)	-	9.70	6.73	0.16	11.66
SHF (same key as target)	4	100.00	100.00	99.81	100
SHF (different key)	4	25.84	15.28	0.27	15.79
SHF	8	13.61	9.8	0.12	16.67
SHF	16	6.58	4.7	0.06	12.92
NP	4	8.37	6.14	0.14	16.02
NP	8	11.88	8.66	0.19	17.89
NP	16	4.17	3.16	0.09	11.88
FFX	4	4.44	2.61	1.22	8.72
FFX	8	2.99	2.23	1.04	7.53
FFX	16	3.68	3.16	1.2	7.41

[3] ——, “Ensemble of key-based models: Defense against black-box adversarial attacks,” in *Proc. of Global Conference on Consumer Electronics*, 2021, pp. 95–98.

[4] H. Kiya, M. AprilPyone, Y. Kinoshita, S. Imaizumi, and S. Shiota, “An overview of compressible and learnable image transformation with secret key and its applications,” *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, e11, 2022.

[5] M. Tanaka, “Learnable image encryption,” in *Proc. of IEEE International Conference on Consumer Electronics-Taiwan*, 2018, pp. 1–2.

[6] K. Madono, M. Tanaka, M. Onishi, and T. Ogawa, “Block-wise scrambled image recognition using adaptation network,” in *Proc. of Workshop on Artificial Intelligence of Things*, 01 2020.

[7] 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.

[8] O. Watanabe, A. Nakazaki, and H. Kiya, “A fast image-scramble method using public-key encryption allowing backward compatibility with jpeg2000,” in *Proc. of International Conference on Image Processing*, vol. 5, 2004, pp. 3435–3438 Vol. 5.

[9] I. Nakamura, Y. Tonomura, and H. Kiya, “Unitary transform-based template protection and its application to l^2-norm minimization problems,” *IEICE TRANSACTIONS on Information and Systems*, vol. E99-D, no. 1, pp. 60–68, 2016.

[10] W. Sirichotedumrong and H. Kiya, “A gan-based image transformation scheme for privacy-preserving deep neural networks,” in *Proc. of European Signal Processing Conference*, 2021, pp. 745–749.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[13] F. Croce and M. Hein, “Minimally distorted adversarial examples with a fast adaptive boundary attack,” in *Proc. of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119, 13–18 Jul 2020, pp. 2196–2205.

[14] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[15] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, “Square attack: A query-efficient black-box adversarial attack via random search,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 484–501.

[16] Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong, “NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks,” in *Proc. of International Conference on Machine Learning*, vol. 97, 09–15 Jun 2019, pp. 3866–3876.

[17] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML’20, 2020.

[18] N. Papernot, P. D. McDaniel, and I. J. Goodfellow, “Transferability in machine learning: from phenomena to black-box attacks using adversarial samples,” *CoRR*, vol. abs/1605.07277, 2016. [Online]. Available: http://arxiv.org/abs/1605.07277

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2014.

[20] Y. Liu, X. Chen, C. Liu, and D. Song, “Delving into transferable adversarial examples and black-box attacks,” in *Proc. of International Conference on Learning Representations*, 2017.

[21] K. Mahmood, R. Mahmood, and M. van Dijk, “On the robustness of vision transformers to adversarial examples,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 7838–7847.

[22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. of International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2015.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. of International Conference on Learning Representations*. OpenReview.net, 2021.