

# Enhanced Security against Adversarial Examples Using a Random Ensemble of Encrypted Vision Transformer Models

<sup>1st</sup> Ryota Iijima

Tokyo Metropolitan University  
Tokyo, Japan  
ijima-ryota@ed.tmu.ac.jp

<sup>2nd</sup> Miki Tanaka

Tokyo Metropolitan University  
Tokyo, Japan  
miikeneko1221@outlook.com

<sup>3rd</sup> Sayaka Shiota

Tokyo Metropolitan University  
Tokyo, Japan  
sayaka@tmu.ac.jp

<sup>4th</sup> 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, which means AEs generated for a source model can fool another black-box model (target model) with a non-trivial probability. In previous studies, it was confirmed that the vision transformer (ViT) is more robust against the property of adversarial transferability than convolutional neural network (CNN) models such as ConvMixer, and moreover encrypted ViT is more robust than ViT without any encryption. In this article, we propose a random ensemble of encrypted ViT models to achieve much more robust models. In experiments, the proposed scheme is verified to be more robust against not only black-box attacks but also white-box ones than convention methods.

**Index Terms**—adversarial example, transferability, ensemble model

## I. INTRODUCTION

Deep neural networks (DNNs) have been developed in various fields, but they have critical problems to be resolved [1]. One of the problems is that DNNs are vulnerable to adversarial examples (AEs), so a trained model is fooled by using AEs. In addition, AEs also have a property, called the transferability of AEs, which means that AEs designed for a model (source model) fool a black-box model (target model) with a non-trivial probability as well as the source model. In this paper, we aim to construct robust models against AEs including the transferability of AEs.

To achieve robust models against AEs, various studies have been reported so far [2]–[8]. In previous studies, it was confirmed that the use of models trained with encrypted images is robust against white-box attacks, but it is not effective under state-of-the-art black-box attacks [4]–[8]. The vision transformer (ViT) was also demonstrated to be more robust against the property of adversarial transferability than convolutional neural network (CNN) models such as ConvMixer, and moreover encrypted ViT is more robust than ViT without any encryption [9].

Because of such a situation, in this paper, we propose a random ensemble of encrypted ViT models to achieve much more robust models. In experiments, the proposed scheme is

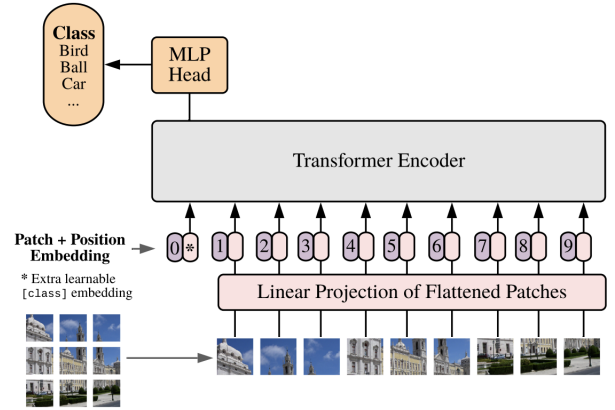


Fig. 1. Architecture of Vision Transformer [10]

verified to be more robust against not only black-box attacks but also white-box ones than convention methods.

## II. RELATED WORK

### A. Vision Transformer

The vision transformer (ViT) [10] is known as a model that provides high performance in classification tasks. Figure 1 shows the architecture of ViT. ViT classifies images according to the following steps.

- 1) Split an image into fixed-size patches, and linearly embed each of them.
- 2) Add position embedding to patch embedding.
- 3) Feed the resulting sequence of vectors to a standard transformer encoder.
- 4) Feed the output of the transformer to a multi-layer perceptron (MLP), and get a result.

ViT is usually used after fine-tuning a pre-trained model. In previous studies, it was shown that fine-tuning by using encrypted images improves the robustness against AEs [9]. In this paper, we also use ViT models fine-tuned with encrypted images as sub-models for a random ensemble.

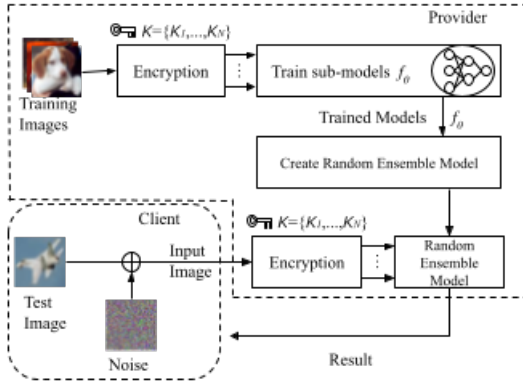


Fig. 2. Framework of proposed scheme.

### B. Adversarial Examples

Depending on the ability of adversaries, there are two types of attacks: white-box attacks [11]–[14] and black-box attacks [15]. The adversaries have complete knowledge of the target model and data information in white-box settings. In contrast, in black-box settings, adversaries can transfer the generated AE to the unknown deployed model based on AE transferability. Furthermore, AEs can be categorized into two types in terms of the goal of adversaries. Target attacks mislead the output of models to a specific class. In contrast, non-targeted attacks aim to mislead models to an incorrect class.

AutoAttack [16] was proposed to evaluate the robustness of defense methods against AEs in an equitable manner. The attack method consists of four parameter-free attack methods: Auto-PGD-cross entropy (APGD-ce), Auto-PGD-target (APGD-t), FAB-target (FAB-t) [17], and Square attack [15]. In this paper, we use APGD-ce and Square attack as a white-box attack and a black-box attack to evaluate an random ensemble, respectively. Both attacks are non-targeted ones.

Adversarial training [11], [18]–[20] is widely known as a defense method against AEs, where AEs are used as training data to improve the robustness against AEs. However, it degrades the performance of models when clean images are input. Defense methods against AEs are expected to meet the following requirements in general.

- No performance degradation even when clean images are input.
- Being robust enough against all attack methods.

## III. PROPOSED METHOD

### A. Overview

Figure 2 shows the framework of the proposed scheme. At first, a provider trains  $N$  sub-models with images encrypted with secret keys  $K = \{K_1, \dots, K_N\}$ . Next, the provider constructs a random ensemble of the sub-models as an image classifier. The provider encrypts a test image with  $K$  to generate  $N$  encrypted test images, and the encrypted images are input to the classifier (a random ensemble of encrypted ViT models) to get an estimate result.

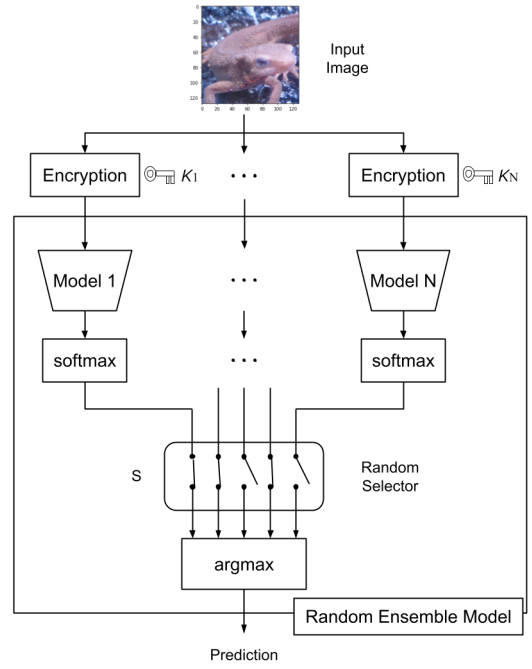


Fig. 3. Random ensemble of encrypted models.

### B. Random ensemble of sub-models

Figure 3 shows the details of a random ensemble of  $N$  encrypted sub-models. Every sub-model is ViT, and a different secret key is assigned to each sub-model for image encryption. In this paper, pixel shuffling is used for image encryption as in [9]. The following steps are carried out to generating encrypted images for pixel shuffling.

- 1) Split an image into non-overlapped blocks with a size of  $M \times M$ , where  $M$  is the same size as the patch size of ViT.
- 2) Flatten each block into a vector.
- 3) Randomly permute pixels in each vector to generate an encrypted vector by using key  $K_i, i = 1, 2, \dots, N$ .
- 4) Rebuild the encrypted vector into the encrypted block.
- 5) Concatenate the encrypted blocks into an encrypted image.

Please note that  $N$  encrypted images are generated from an image by using  $N$  keys, and any clients do not know the keys. In the proposed method,  $S$  outputs are randomly selected from  $N$  outputs of sub-models where  $3 \leq S \leq N$ . The final outputs are determined by the average of  $S$  outputs.

## IV. EXPERIMENT

### A. Experimental Setup

Experiments were conducted on the CIFAR-10 dataset. The dataset, which consists of 60,000 images with size  $32 \times 32 \times 3$ , was divided into 50,000 and 10,000 images for fine tuning and testing, respectively. All images were resized to  $224 \times 224 \times 3$  to fit the input to ViT and scaled to  $[0, 1]$  as a range of the values. We used finetuned ViT models with a patch size of

TABLE I  
COMPARISON OF ENSEMBLE MODEL AND RANDOM ENSEMBLE MODEL  
WITH ASR (ALL KEYS ARE KNOWN TO AN ATTACKER)

Source Model	Attack Method	
	APGD-ce	Square
Ensemble	100.00	98.43
Random Ensemble (Proposed)	99.90	<b>22.21</b>

TABLE II  
ASR OF RANDOM ENSEMBLE MODEL (SOME KEYS ARE KNOWN TO AN  
ATTACKER)

Source Model		Attack Method
Model	# leaked keys	APGD-ce
Ensemble	4	100.00
	2	97.25
	1	2.24
	0	0.04
Random Ensemble (Proposed)	4	99.90
	2	71.74
	1	2.73
	0	0.12

$P = 16$  where ViT was pre-trained with ImageNet-21k [10]. ImageNet-21k is a dataset consisting of 21,000 classes with a total of 1,400 million patches, which were resized to image size  $224 \times 224 \times 3$  when pre-training ViT. For fine tuning, a learning rate of  $lr = 0.03$  was set, and we ran 5,000 epochs. For model encryption, a block size of  $M = 16$  was used as well as the patch size of ViT. The robustness models were evaluated by using two attack methods, APGD-ce, which is included in AutoAttack [16] as a white-box attack, and Square attack [15], which is also included in AutoAttack as a black-box attack. The Attack Success Rate (ASR) was used as an evaluation metric.

### B. Experimental Result

First, we compared the proposed random ensemble models with ensemble models (no random selection) under the use of APGD-ce and Square attack (see Table I), where both models consisted of  $N = 4$  sub-models encrypted by using different keys. From Table I, the proposed ensemble outperformed the conventional one under Square, but both ensembles had almost the same ASR values under APGD-ce where it was assumed that an attacker knew all keys of sub-models.

In Table 2, Next, the ASR values of random ensemble models against APGD-ce were evaluated when the number of keys known to an attacker was varied. As show in Table II, APGD-ce attacks failed when an attacker did not know any keys or knows only one key.

## V. CONCLUSION

The use of encrypted models was known to be effective against white-box attacks if secret keys are not open. In this paper, we proposed an novel method with encrypted models, which is carried out on the basis of an random ensemble of encrypted ViT models, so that the robustness of models

is enhanced against black-box attacks in addition to against white-box attacks when disclosing a few keys.

## ACKNOWLEDGMENTS

This study was partially supported by JSPS KAKENHI (Grant Number JP21H01327) and JST CREST (Grant Number JPMJCR20D3).

## REFERENCES

- [1] H. Kiya, A. P. M. Maung, 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, 2022. [Online]. Available: <http://dx.doi.org/10.1561/116.000000048>
- [2] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu, "Improving adversarial robustness via promoting ensemble diversity," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 4970–4979. [Online]. Available: <https://proceedings.mlr.press/v97/pang19a.html>
- [3] H. Yang, J. Zhang, H. Dong, N. Inkawhich, A. Gardner, A. Touchet, W. Wilkes, H. Berry, and H. Li, "Dverge: Diversifying vulnerabilities for enhanced robust generation of ensembles," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.
- [4] A. MaungMaung 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.
- [5] —, "Privacy-preserving image classification using an isotropic network," *IEEE MultiMedia*, vol. 29, no. 2, pp. 23–33, 2022.
- [6] —, "Encryption inspired adversarial defense for visual classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1681–1685.
- [7] —, "A protection method of trained cnn model using feature maps transformed with secret key from unauthorized access," *APSIPA Transactions on Signal and Information Processing*, vol. 10, p. e10, 2021.
- [8] —, "Ensemble of key-based models: Defense against black-box adversarial attacks," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, pp. 95–98.
- [9] M. Tanaka, I. Echizen, and H. Kiya, "On the transferability of adversarial examples between encrypted models," in *2022 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2022, pp. 1–4. [Online]. Available: <https://doi.org/10.1109/ISPACS57703.2022.10082844>
- [10] 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 *International Conference on Learning Representations*, 2021.
- [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*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1412.6572>
- [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*. OpenReview.net, 2018. [Online]. Available: <https://openreview.net/forum?id=rJzIBfZAb>
- [13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.
- [14] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2574–2582.
- [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] 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. JMLR.org, 2020.
- [17] —, “Minimally distorted adversarial examples with a fast adaptive boundary attack,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML’20. JMLR.org, 2020.
- [18] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” in *International Conference on Learning Representations*, 2017. [Online]. Available: <https://openreview.net/forum?id=BJm4T4Kgx>
- [19] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 7472–7482. [Online]. Available: <http://proceedings.mlr.press/v97/zhang19p.html>
- [20] Y. Carmon, A. Raghuathan, L. Schmidt, J. C. Duchi, and P. S. Liang, “Unlabeled data improves adversarial robustness,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.