Adversarial Test on Learnable Image Encryption

MaungMaung AprilPyone, Warit Sirichotedumrong and Hitoshi Kiya

Department of Computer Science, Graduate School of System Design

Tokyo Metropolitan University

Asahigaoka, Hino-shi, Tokyo, 191-0065, Japan

{april-pyone-maung-maung@ed.,warit-sirichotedumrong@ed.,kiya@}tmu.ac.jp

Abstract—Data for deep learning should be protected for privacy preserving. Researchers have come up with the notion of learnable image encryption to satisfy the requirement. However, existing privacy preserving approaches have never considered the threat of adversarial attacks. In this paper, we ran an adversarial test on learnable image encryption in five different scenarios. The results show different behaviors of the network in the variable key scenarios and suggest learnable image encryption provides certain level of adversarial robustness.

Index Terms—learnable image encryption, adversarial robustness

I. Introduction

Deep learning has brought major breakthroughs in computer vision as well as other fields [1]. There is no doubt that it is due to better algorithms, bigger data and faster computing resources. With large amount of data, deep learning is often carried out in the cloud environments where privacy issues are generated. To securely transmit images through an untrusted channel, researchers have proposed Encryption-then-Compression (EtC) systems such as [2]–[4]. Nevertheless, most of the traditional image encryption schemes are not compatible with deep learning. Recently, a learnable image encryption method where a neural network can learn encrypted images [5] and pixel-based image encryption designed for deep neural networks [6] were proposed.

Although privacy issues have been addressed, the security of deep learning has never been considered in the context of learnable image encryption. Adversarially robust models are extremely desired for privacy-preserving schemes because the protected data are usually sensitive. Unexpected misclifications due to adversarial attacks will lead to severe damage to applications such as medical analysis, surveillance, etc. Therefore, security is quintessential in privacy-preserving deep neural networks. In the literature, it has been shown that state-of-the-art neural networks are vulnerable towards adversarial examples [7], [8]. Adversarial examples can be generated by optimization techniques to maximize the loss such as [9]–[11]. Since then, deep learning has got significant amount of attention towards adversarial robustness [12].

Our hypothesis is that encryption for privacy protection can give us somewhat adversarial robustness. In this paper, we run a test on a recent learnable encryption method [5] where input images are encrypted and sent to the network with an adaptation layer. From our experiments, we confirm that learnable image encryption has somewhat resistance to
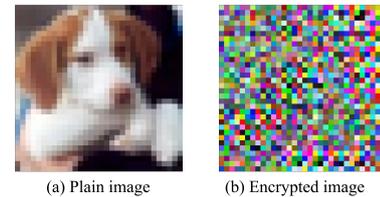

(a) Plain image   (b) Encrypted image
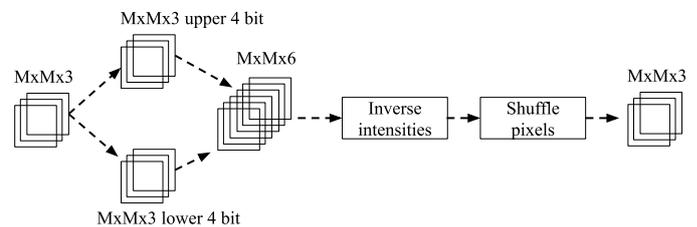
Fig. 1. An example of encrypted image.

Fig. 2. Encryption process of a block by Tanaka's method [5].

adversarial examples. Our contribution in this paper is that we conducted a test to raise a fundamental need of adversarial robustness in the privacy-preserving network. We are also the first to consider adversarial perspective in learnable image encryption.

II. Preliminaries

A. Learnable Image Encryption

Fig. 1 illustrates an example of encrypted image by Tanaka's encryption. There are 2 parts in Tanaka's work: image encryption and adaptation layer [5].

The 8-bit pixel values in MxM blocks are separated into upper and lower 4-bit to form the 6-channel blocks. The intensities of pixel values are randomly reversed and shuffled. The 6-channel blocks are reformed to 3-channel blocks. For simplicity, the encryption process of one block is depicted in Fig. 2.

The adaptation comprises of the first convolution layer (MxM kernel and MxM stride), several network-in-network style layers and sub-pixel convolution (pixel shuffle). After the adaptation network, any network can be followed. Fig. 3 describes the diagram of Tanaka's adaptation network.

B. Adversarial Examples

Adversarial examples are perceptually same images as the original ones that are carefully designed to make the neural
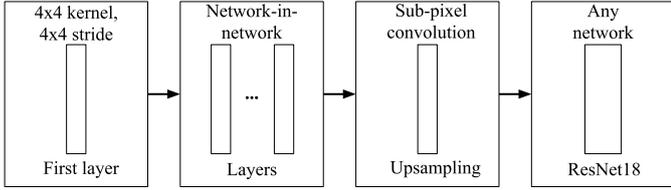
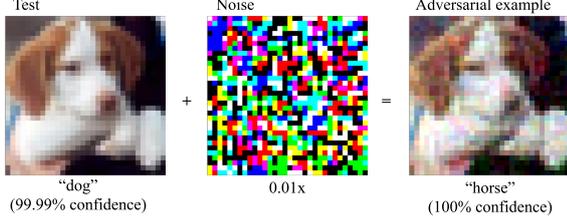Fig. 3. Diagram of adaptation network by Tanaka's method [5].

Test
"dog"
(99.99% confidence)

Noise
0.01x

Adversarial example
"horse"
(100% confidence)

Fig. 4. A sample of adversarial example generated by PGD.

Fig. 5. The flow diagram of the experiment ("Adv. Training" refers to "Adversarial Training").

network misclassify with high confidence. Fig. 4 shows an example of adversarial example where the network classifies "dog" as "horse" with 100% confidence.

There are many different ways of crafting adversarial examples. The popular and computationally efficient one is known as Fast Gradient Sign Method (FGSM) [9]. In this work, we consider a stronger adversary (i.e., multi-step FGSM) known as projected gradient descent (PGD) [10]:

$$x^{t+1} = clip_\epsilon(x^t + \alpha \, sign(\nabla_x L(\theta, x, y))), \qquad (1)$$

where ϵ is allowable perturbation size, x^t is adversarial example on t^{th} PGD iteration, α is step size and $\nabla_x L(\theta, x, y)$ represents the gradients of the loss function. Instead of multiplying the sign of the gradients with ϵ directly, PGD only adds some step size α to the image in each iteration. PGD also projects the perturbation back to the max-norm in each step (i.e., $clip_\epsilon(.)$).

C. Adversarial Training

Adversarial training is to train a network to be robust against adversarial examples. There are many types of adversarial defense. Madry et al. show that training against PGD leads to robustness against other first-order adversaries [12]. Therefore, we focus on PGD training in this work for experimental purposes.

III. EXPERIMENTS

The flow diagram of the experiment is presented in Fig. 5. First, we encrypted the data and used the encrypted images to train the network by reusing the encryption code released by the author in GitHub [13]. For the testing, We generated adversarial noise by PGD, added the noise to the test images and then, encrypted them.

The experiment setup is described as follows. We used CIFAR10 [14] dataset with batch size of 128. Then, we implemented Tanaka's adaptation layer on deep residual network (ResNet18) [15] on PyTorch platform. The network
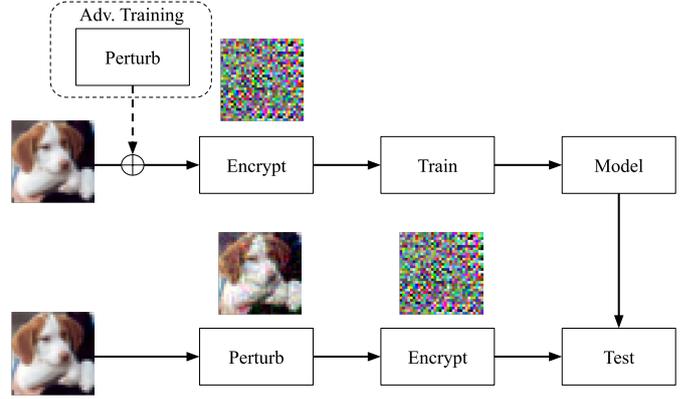
was trained for 100 epochs with stochastic gradient descent optimization on learning rate of 0.1. We reduced the learning rate to 0.01 after running for 40 epochs. All the images in the dataset for all cases were in the range of $[0, 1]$ with live augmentation (random cropping and random horizontal flip). However, there was no prior normalization. For the adversarial testing and training, we used PGD [10] with $\epsilon = 0.1$, $\alpha = 0.01$ for 20 iterations.

Adversarial noise was generated iteratively as described in (1). We used a slightly higher noise level (i.e., $\epsilon = 0.1$) to stress the experiment. The PGD process with the same settings was employed in both adversarial testing and training.

We trained the network in 5 different scenarios and tested against adversarial examples generated by PGD.

1) Plain: ResNet18 trained on clean images
2) Encrypted: ResNet18 with Tanaka's adaptation layer trained on encrypted images (one key for both training and testing)
3) Encrypted-Adv: Adversarial training for encrypted images using the same key
4) Encrypted-DK: Same model as **Encrypted** except each batch of the dataset was encrypted using a different random key
5) Encrypted-Adv-DK: Adversarial training for encrypted images using different random keys

IV. RESULTS

Table I summarizes the results of the experiments. The error is the accuracy of misclassification (the lower the better). We recorded the errors during training, testing and adversarial testing. The results suggest the followings:

1) The standard model trained with plain images are completely vulnerable towards adversarial examples (i.e., 100% misclassification).
2) Image encryption provides certain degree of robustness against adversarial examples. The error rate of 0.269 is extremely good for adversarial robustness. Even the state-of-the-art adversarially trained model had the approximate error rate of 0.530 for PGD attack with 20

Model	Error		
	Train	Test	Adversarial
Plain	0.006	0.137	1.000
Encrypted	**0.011**	**0.166**	**0.269**
Encrypted-Adv[†]	0.093	0.830	0.040
Encrypted-DK[‡]	0.431	0.506	0.792
Encrypted-Adv[†]-DK[‡]	0.657	0.617	0.690

[†]Adversarial, [‡]Diffferent keys.

steps [16]. This confirms our hypothesis that the encryption can give certain level of adversarial robustness.

3) Traditional adversarial training is not suitable for learnable image encryption. While doing adversarial training with the same key, the network became biased towards adversarial examples. The network could not generalize the clean examples. Therefore, the network performed poor on the test dataset.

4) The neural network requires uniform encryption for both train and test data. Different keys transform the data into different distributions. Therefore, when using different keys for encryption, the network did not learn.

5) Since the network assumes encryption with one key, adversarial training with different keys is not relevant. We carried out the test for experimental purposes only.

V. CONCLUSIONS

Learnable image encryption provides good adversarial robustness. The same key is necessary to uniformly encrypt the images. In addition, we can also conclude that the standard adversarial training is not suitable for encrypted images. We aim to achieve secure and private models with reasonable accuracy as our future work.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[2] 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.

[3] 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.

[4] K. Kurihara, S. Shiota, and H. Kiya, "An encryption-then-compression system for jpeg standard," in *2015 Picture Coding Symposium (PCS)*. IEEE, 2015, pp. 119–123.

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

[6] 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

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[8] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[10] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[11] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[13] M. Tanaka, "Learnable image encryption," https://github.com/mastnk/ICCE-TW2018.

[14] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] MadryLab, "A challenge to explore adversarial robustness of neural networks on cifar10," https://github.com/MadryLab/cifar10_challenge.