Black-Box Adversarial Attack against Deep Neural Network Classifier Utilizing Quantized Probability Output

Haruki Masuda¹, Tsunato Nakai^{1,2}, Kota Yoshida¹, Takaya Kubota¹, Mitsuru Shiozaki¹ and Takeshi Fujino¹

¹ Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu-shi, Shiga 525-8577, Japan
E-mail: ri0068rr@ed.ritsumei.ac.jp, fujino@se.ritsumei.ac.jp

² Mitsubishi Electric Corp., 5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, Japan

Abstract

Deep neural networks (DNNs) are vulnerable to well-designed input samples, known as adversarial examples. In particular, an attack involving the generation of adversarial examples is called a black-box attack when an adversary attacks without any internal knowledge of the target network. In a simple black-box attack, adversarial perturbations are selected on the basis of changes in output probability when the input to the DNN is slightly changed. Output probability quantization has been proposed as a countermeasure against the simple black-box attack.

In this work, we quantitatively evaluate the effectiveness of this protection method by using the image degradation index and propose a new black-box attack that can overcome the output probability quantization. We conducted experiments to generate adversarial examples using the MNIST public dataset. In the conventional method, if the fourth digit after the decimal point of the output probability is truncated, perturbations that can easily be recognized by humans appear in the adversarial example, and the attack ability decreases. With the new attack method, we find that adversarial examples can be generated with a sufficiently small degradation even if the output probability is truncated after the second decimal place. This demonstrates that the output probability quantization countermeasure against the simple black-box attack is not effective.

1. Introduction

Deep neural networks (DNNs) are increasingly being utilized in image recognition, self-driving, and person detection systems used with security cameras. To apply DNNs to fields related to human safety and security, it is critical to consider malicious attacks. One such attack is an adversarial example [1] that induces the misclassification of the DNN's classifiers with a small perturbation that cannot be recognized. In general, white-box attacks, which exploit internal knowledge of target models (such as

network configuration and weight parameters), are used to create adversarial examples. However, in a black-box attack scenario, an attack is performed without target models. This type of attack is more threatening because model protection is not effective. In simple black-box attacks [2], perturbations are selected on the basis of changes in output probability when the input to the DNN is slightly changed. Simple black-box attacks are easy to implement and exploit.

Output probability quantization has been proposed as a countermeasure against simple black-box attacks. Senzaki et al. [3] proposed the quantization of output probability (reduction in the number of output digits) on an inference system as a countermeasure against a simple black-box attack on a DNN in an image classification task, and Bhagoji et al. [4] reported that the quantization of the output probability can reduce the success rate of the attack. In this work, we evaluate the effect of output probability quantization in a simple black-box attack by using the image degradation evaluation index and propose a new black-box attack that can overcome the probability quantization.

2. Conventional and Proposed Methods

2.1 Conventional method: Simple black-box attack

The conventional method (simple black-box attack) is described by Algorithm 1 and is illustrated in Fig.1(a). The perturbation applied to the input is given as $+\alpha$, $-\alpha$ independently for each pixel. By setting α to a sufficiently small value, adversaries can generate adversarial examples with a sufficiently small perturbation. The effect of quantizing the output probability is shown in Fig.1(b). If α is very small, the output probability will not change. Therefore, it is necessary to increase α , but the generated adversarial examples will be greatly degraded compared with the original image.

Figure 1: Effect of quantization against simple black-box attack

Figure 2: Proposed attack method against the countermeasure utilizing output probability quantization

Algorithm 1: Simple black-box attack [2]

Input: Image X , True class label y_{true} ,
Number of image pixels, N
Perturbation parameter α
Classifier ϕ

Output: Adversarial example X^{adv}

```

1:  $X^{adv} = X$ 
2: while  $\text{argmax}(\phi(X^{adv})) = y_{true}$ 
3:    $x = X^{adv}$ 
4:   for  $n=0$  to  $N-1$  do
5:      $x' = x$ 
6:      $x'_n = x_n + \alpha$ 
7:      $p^+ = \text{probability of } y_{true} \text{ in } \phi(x')$ 
8:      $x' = x$ 
9:      $x'_n = x_n - \alpha$ 
10:     $p^- = \text{probability of } y_{true} \text{ in } \phi(x')$ 
11:    if  $p^+ > p^-$ 
12:       $X_n^{adv} = X_n^{adv} - \alpha$ 
13:    else if  $p^+ < p^-$ 
14:       $X_n^{adv} = X_n^{adv} + \alpha$ 
15:    else
16:       $X_n^{adv} = X_n^{adv} + 0$ 
17:    end if
18:     $X^{adv} = \text{clip}_{[0,1]}(X^{adv})$ 
19:  end for
20: end while
21: return  $X^{adv}$ 

```

2.2 Proposed method: Advanced black-box attack

In section 2.1, we showed that if the output probability is quantized, resistance against the simple black-box attack is provided. Therefore, we propose an advanced black-box attack that can generate adversarial examples even if the output probability is quantized. The proposed attack method is described by Algorithm 2 and is illustrated in Fig.2. In the conventional attack method, the parameter α is used as a value for determining both the direction (sign) and the amount of perturbation. In contrast, we introduce a new parameter β for determining the direction of the perturbation and use the parameter α for determining the amount of perturbation. The adversary can conjecture the direction of the gradient from a sufficiently large β even if the output probability is quantized. The image degradation can be decreased because the perturbation increment is suppressed by the small α .

3. Experiments and Results

3.1 Simulation model and sample images

We use the peak signal-to-noise ratio (PSNR) to evaluate the image degradation of the adversarial examples. The PSNR is an index commonly used to determine the degree of image degradation. A larger PSNR means less degradation from the original image. The PSNR is defined as

$$MSE = \frac{1}{N} \sum_{n=0}^{N-1} (X_n^{adv} - X_n)^2 \quad (1)$$

$$PSNR = 20 \log_{10} (MAX / \sqrt{MSE}) \quad (2)$$

Algorithm 2: Advanced simple black-box attack

```

1:    $X^{adv} = X$ 
2:   while  $\text{argmax}(\phi(X^{adv})) = y_{true}$ 
3:      $x = X^{adv}$ 
4:     for  $n$  to  $N$  do
5:        $x' = x$ 
6:        $x'_n = x_n + \beta$ 
7:        $p^+ = \text{probability of } y_{true} \text{ in } \phi(x')$ 
8:        $x' = x$ 
9:        $x'_n = x_n - \beta$ 
10:       $p^- = \text{probability of } y_{true} \text{ in } \phi(x')$ 
11:      if  $p^+ > p^-$ 
12:         $X_n^{adv} = X_n^{adv} - \alpha$ 
13:      else if  $p^+ < p^-$ 
14:         $X_n^{adv} = X_n^{adv} + \alpha$ 
15:      else
16:         $X_n^{adv} = X_n^{adv} + 0$ 
17:      end if
18:       $X^{adv} = \text{clip}_{[0,1]}(X^{adv})$ 
19:    end for
20:  end while
21:  return  $X^{adv}$ 

```

where X_n is the value of one pixel of the original image X , N represents all pixels of the original image X and MAX is the maximum possible pixel value of the original image. Generally, if the PSNR exceeds 30dB, the difference between the original and processed images is invisible to the human eye.

For the evaluation, we use MNIST [5], which is a public dataset of handwritten digits for image classification tasks. We use 60,000 datasets for training the target model and another 10,000 for testing. As shown in Table 1, the target DNN model consists of three layers: an input layer, a fully connected layer and an output layer. The test accuracy of the target DNN model is 97.27%. The generation and evaluation of adversarial examples are performed on a total of ten images randomly selected from each of the classes 0–9 that are successfully classified from the 10,000 test datasets. Figure 3 shows each predicted class and its output probability for the selected samples.

3.2 PSNR evaluation of each method

First, we generated adversarial examples under two conditions: simple black-box attacks with and without quantization. The parameter α is selected to be the minimum value for a successful attack. The range of

Table 1: Architecture of classifier for MNIST

Layer type (data shape)
Input (28, 28)
Flatten (784)
Fully connected + ReLU (64)
Softmax (10)

True label	0	1	2	3	4
Sample image
Predicted label	0	1	2	3	4
Probability	0.9999418	0.9999999	0.989527	0.9998337	0.9822557
True label	5	6	7	8	9
Sample image
Predicted label	5	6	7	8	9
Probability	0.9999045	0.9999976	0.9999998	0.9992405	0.9999991

Figure 3: Samples used to generate and evaluate adversarial examples

possible α values ($\alpha \in \{1/255, 2/255, \dots, 255/255\}$) is in accordance with the 8bit resolution of the luminance value of the image. Figures 4,5 show the adversarial examples and the perturbations generated for the image of class 7 under each quantization condition. The perturbations $X^{adv} - X$ are magnified for easy visualization and the magnification is shown above the figure of the perturbations.

The results of the conventional method are shown in Fig.4. When the number of digits of the output probability is four, the PSNR of the image of class 7 falls below 30dB. Furthermore, when the number of digits is less than three, the PSNR falls below 10dB. These results demonstrate that the adversarial examples generated with the simple black-box attack are greatly degraded by quantizing the output probability to four digits or less.

Similarly, we evaluate the proposed advanced black-box attack. The parameter β is set to 1.0, and α is selected to be a minimum value for a successful attack. Figure 5 shows the adversarial examples generated for each image of class 7 using the proposed method. The PSNRs of the adversarial examples exceed 30 dB even if the output probability is quantized.

Figure 6 shows the PSNRs of the adversarial examples generated from the samples shown in Fig.3 by the conventional and proposed methods. When the output probability is quantized to three digits or less, the PSNR of the adversarial examples generated by the conventional method varies. This is affected by the difference in probability between the top-1 and top-2 predictions of the neural network. If this difference is large, the PSNR decreases because the adversarial examples cannot be created unless the perturbation is sufficiently large.

Figure 4: Adversarial examples generated under each condition for the image of class 7 using the conventional attack method

Figure 5: Adversarial examples generated under each condition for the image of class 7 using the proposed attack method

Figure 6: PSNRs of adversarial examples generated using the conventional and proposed attack methods

4. Conclusion

In this paper, we evaluated the image degradation of adversarial examples using a simple black-box attack when the output probability of the DNN model is quantized. The PSNR was used to evaluate the degradation. We also evaluated a countermeasure against the simple black-box attack through experiments using the MNIST public dataset for an image recognition task. We found that by quantizing the fourth and subsequent decimal digits of the output probability, the generated perturbations greatly degrade the adversarial examples.

We then proposed an advanced black-box attack that can attack even if the output probability is quantized. In the proposed attack method, a large perturbation value for determining the direction of the perturbation and a small perturbation value for updating the attack image are applied as separate parameters. We found that even if the output probability is quantized to the first decimal place, the proposed method can generate adversarial examples with

sufficiently small degradation.

These results demonstrate that even if one digit of the output probability of the target class is output, it is possible to generate adversarial examples with little degradation using the advanced black-box attack proposed in this paper. Therefore, adversarial examples generated by the proposed method are a major threat in edge AI, where it is difficult to apply other countermeasures such as limiting the number of inference queries. Countermeasures other than output quantization should be considered.

Acknowledgment

This work was supported by JST-Mirai Program Grant Number JPMJMI19B6, Japan.

References

- [1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus: Intriguing properties of neural networks, arXiv:1312.6199, 2013.
- [2] N. Narodytska and S. P. Kasiviswanathan: Simple black-box adversarial perturbations for deep networks, arXiv:1612.06299, 2016.
- [3] Y. Senzaki, S. Ohata and K. Matsuura: Black-box attack for deep neural networks with efficient generation of adversarial examples, Symposium on Cryptography and Information Security, 2018.
- [4] A. N. Bhagoji, W. He, B. Li and D. Song: Practical black-box attacks on deep neural networks using efficient query mechanisms, European Conference on Computer Vision, 2018.
- [5] L. Yann and C. Corinna: The MNIST database of handwritten digits, <http://yann.lecun.com/exdb/mnist/>.