Image hashing algorithm to defend FGSM attacks on Neural Network

Junyaup Kim
Department of Computer Science and Engineering
Sungkyunkwan University
Suwon, South Korea
yaup21c@g.skku.edu

Siho Han
Department of Applied Data Science
Sungkyunkwan University
Suwon, South Korea
siho.han@g.skku.edu

Simon S. Woo
Department of Computer Science and Engineering
Sungkyunkwan University
Suwon, South Korea
swoo@g.skku.edu

ABSTRACT

In this research, we present a performance evaluation of existing image hashing algorithms on defending deep learning models against adversarial attacks as an initial work to developing a new, time-efficient image hashing algorithm. More specifically, we aim to set a maximum time complexity of $O(N^2)$ as a constraint to the algorithm, such that it can be used in time-critical systems and/or low computing resources systems. Upon experimenting with existing image hashing algorithms, we conclude that the wavelet hashing algorithm achieves the highest accuracy (75%) when detecting images generated from Neural Networks attacked by the FGSM, with a time complexity of $O(N)$.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; Neural networks; *Image manipulation.*

KEYWORDS

Deep learning, Adversarial Attack, Image hashing, Image cryptography

1 INTRODUCTION

Deep learning-based algorithms are known to show superior performance when tackling various computer vision asks, such as image recognition, object detection, and segmentation. However, Convolutional Neural Network(CNN)-based models like CNN image classifiers are susceptible to adversarial perturbations, which are designed to cause critical faults in deep neural network models by adding synthetic noises, imperceptible to human observers, to input images. Moreover, the deployment of modern systems leveraging deep learning is under time and resource constraints, precluding them from using additional modules that deal with adversarial attacks. In this paper, we explore different image hashing algorithms used for blocking adversarial attacks and show that the wavelet hashing algorithm-based thresholding method achieves the highest accuracy. Note that in this paper, we solely focus on untargeted attacks on a specific model, such that this work serves as a basis for future research.

2 RELATED

Given an input x with a ground truth label y and a classifier $f_\theta(\cdot)$ with a set of parameters, an adversarial sample x' is defined as the generated output close to x in terms of a measurable distance, such as the L_p norm ($0 \leqq p \leqq \infty$). There are two kinds of adversarial attack scenarios. One is untargeted attack that distorts the input image an unintended prediction, $f_\theta(x' \neq y)$ and the other is the targeted attack, $f_\theta(x' = y^*)$, for a specific y^* class, which is different from y. Goodfellow *et al.* proposed FGSM(Fast Gradient Sine Method) that applies a first-order approximation of the loss function to construct adversarial samples [1]. In addition, optimization-based methods have also been proposed to create adversarial perturbations for targeted attacks.

3 APPROACH

Our approach assumes that can utilize training dataset of model. To fit on constraints of our research, Image hash function is needed to minimize training duration and classification time. we propose the wavelet hashing algorithm-based thresholding [7] to screen images at a low time complexity.

4 EXPERIMENTS

Our work examines simple perturbations on a basic MNIST dataset [4] for the Le-net model [3].

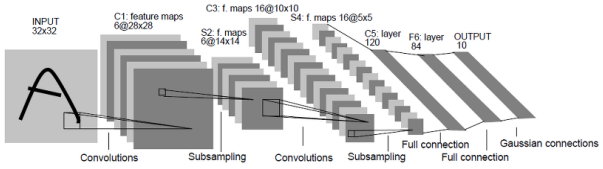

Figure 1: Le-net architecture

We picked a simple model as shown in figure 1, because despite the presence of non-linear activations in CNN models, an FGSM attack assumes that the feature map linearity increases with the model depth. This implies that deeper networks are more vulnerable to FGSM attacks than shallow ones. In an FGSM attack scenario, we set the value of epsilon ranging from zero to 0.3 with an increasing step size of 0.05. Then, we set a threshold based on the output of the wavelet hashing algorithm and test it on unseen data. We generated 26,842 perturbed observations and used 28,000 observations from a real image dataset, giving an approximately equal class ratio. 10% of the observations are used for testing and 20% are used for validation. The remaining data are used to calculate the mean and standard deviation from the resulting image hash values. For the machine configuration, we used Intel(R) Core(TM) i7-9700 CPU 3.00GHz with 16.0GB RAM. For our implementation, we used Python v3.6.8 with the ImageHash v4.0 package.

5 RESULT

In our experiment, the accuracy of $f_\theta(\cdot)$ has been successfully reduced by the FGSM attack, as shown in figure 2. Here, we can

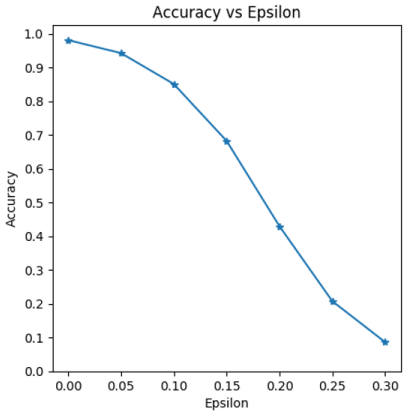

Figure 2: Accuracy vs. epsilon for the model $f_\theta(\cdot)$

observe that Le-net achieves an accuracy of 98% on the MNIST dataset when there is no perturbation. However, the accuracy drops to 10%, which is equivalent to random guessing, when increasing the epsilon value. Epsilon value is the size of perturbation that implemented on input data. In figure 3, we can observe that the

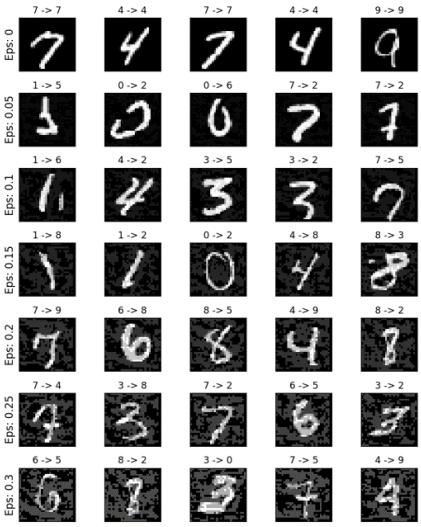

Figure 3: Perturbed output based on Epsilon value

image has more adversarial noises when we increase the epsilon value to generate more severe perturbations. The value shown on the left of each number image is the real value and that shown on the right is the classified output. In figure 4, we calculated the standard deviation and mean values of image hashing output on the training dataset to extract the best threshold.In training dataset, We picked the hashing algorithm with the largest difference between

Algorithm	Perturbed		Genuine	
	Std	Mean	Std	Mean
Average Hash	6.7945e+18	7.9494e+18	3.1978e+17	8.1143e+16
Perceptual Hash	2.6439e+18	1.4565e+19	8.7806e+17	1.0982e+19
difference hashing	5.0728e+18	8.8168e+18	3.1392e+18	5.5686e+18
Wavelet hashing	**6.7113e+18**	**8.4549e+18**	**2.4902e+18**	**2.2742e+18**

Figure 4: The output of each hashing algorithm based on the output of genuine and perturbed image

the perturbed mean and genuine mean, and with a small standard deviation with both classes so that the distribution with each other is easily separable. With the validation set, we checked the difference between the training set mean and standard deviation value. After that, the algorithm is evaluated by Equation 1.

$$min(|\frac{1}{(mean(P) - mean(G))}| + std(P) + std(G)) \quad (1)$$

In equation 1, P denotes the perturbed dataset and G denotes the genuine dataset. This optimization equation implies that the mean values of each distribution should be large, whereas the standard deviation values should be small. By using Equation 1, we can find the optimal algorithm that distinguishes the perturbed images from real images.

$$\lambda = mean + 2std \quad (2)$$

We set the threshold of the image hash values by Equation 2. With Equation 1, we can guarantee that each dataset's distributions have long distance from each other. In this experiment, we only used the validation dataset to generate the mean and standard deviation values. By using the wavelet hashing algorithm with Equation 2 threshold, a test accuracy of 76.63% was achievend.

6 EVALUATION RESULTS AND LIMITATIONS

Even though the wavelet algorithm is not a learnable algorithm, the output shows that the threshold has some effect on discriminating perturbed images. There are artifact patterns on the output of adversarial networks[5]. We can assume that such patterns can be calculated by image hashing algorithms.

7 FUTURE WORK

Our research will focus on developing an image hashing algorithm that can generate hash values from the training dataset and screen for perturbed images. Deeper models, such as ResNet [2] or VGGNet [6] will be explored, and additional adversarial attack scenarios will be added to further verify our approach.

8 CONCLUSION

In this paper, we evaluated the performances of various image hashing algorithms under conditions to generate the optimal threshold between perturbed and genuine images. We showed that the wavelet hashing algorithm can detect robust perturbations on images with an accuracy of 76.63% and a time complexity of $O(N)$.

REFERENCES

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. arXiv:arXiv:1412.6572

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:arXiv:1512.03385
[3] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2323. https://doi.org/10.1109/5.726791
[4] Yann LeCun and Corinna Cortes. 2010. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/. (2010). http://yann.lecun.com/exdb/mnist/
[5] Augustus Odena, Vincent Dumoulin, and Chris Olah. 2016. Deconvolution and Checkerboard Artifacts. *Distill* (2016). https://doi.org/10.23915/distill.00003
[6] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:arXiv:1409.1556
[7] S. P. Singh and G. Bhatnagar. 2017. A robust image hashing based on discrete wavelet transform. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. 440–444. https://doi.org/10.1109/ICSIPA.2017.8120651