# Defensive denoising methods against adversarial attack

Sungyoon Lee
Seoul National University
Seoul, Korea
goman1934@snu.ac.kr

Saerom Park
Seoul National University
Seoul, Korea
psr6275@snu.ac.kr

Jaewook Lee
Seoul National University
Seoul, Korea
jaewook@snu.ac.kr

## ABSTRACT

Deep neural networks are highly vulnerable to adversarial examples. An adversarial example is an image with a small perturbation designed to make networks misclassify it. In this paper, we propose two defensive methods. First, we apply denoising methods using ROF model and NL-means model before classification to remove adversarial noise. Second, we perturb images in certain directions to escape from the adversarial area. Experiments on the universal adversarial perturbations show that proposed methods can remove adversarial noise and perform better classification.

## CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Security and privacy** → *Systems security*;

## KEYWORDS

Adversarial attack; deep neural network; defense; image restoration

## 1 INTRODUCTION

Recently, deep neural networks show powerful results in the image classification tasks [3, 9, 10]. However, Szegedy et al. has shown that small noises that are hard to detect with human eyes can fool the networks [11]. The noise is called adversarial noise and the image with this noise is called adversarial example. This security problem can seriously hinder the application of deep learning to safety-critical systems such as self-driving vehicles, medical image analysis and facial recognition.

There are several methods to generate adversarial examples such as fast gradient sign method (FGS) [2], iterative FGS (IFGS) [4], and universal adversarial perturbations (UAP) [6]. Whereas FGS and IFGS models create adversarial examples dependent on the original input images as well as the target trained networks, UAP model aims to obtain one universal noise that is effective regardless of the input images. The key idea of the UAP is that the noise with 'directional information' can fool the classifier better than random noise. Surprisingly, this effect does not depend on particular images or deep network models. The algorithm involves iterative use of the DeepFool [5] and projection. DeepFool algorithm is based on that the noise pushing data points near to the decision boundaries can degrade the performance of the classifier and the magnitude of this noise can be considered to measure the classifier's robustness. The goal of the algorithm is not to find the nearest decision boundary, but to find a sufficiently close boundary. It is therefore safe to assume that boundaries can be approximated as linear hyperplanes.

In this paper, we propose new methods to defend against adversarial attacks by removing noise of the perturbed images and applying additional defensive noise. We implement ROF model [7] and NL-means model [1] to remove such perturbations and show the result of the methods according to denoising degree. We also suggest to construct a noise that can improve the performance of the classifier by moving the input image to more probable region. The construction scheme is inspired by the idea of UAP [6].

## 2 PROPOSED METHODS

In order to defend against attack using adversarial noise, we first consider denoising before classification model. We use two noise removal algorithms (ROF and NL-means). The UAP contains the directional and metric information of decision boundary of the target classifier and this perturbation is not concentrated in a local area of the image. This kind of noise can be easily removed with denoising models. The ROF and NL-means models effectively forget this information without degrading the classification performance on the undamaged images.

An image can be represented as a vector in the high-dimensi onal space, $v \in \mathbf{R}^N$. Thus, an imperceptible noise, $\delta$ as a vector with element-wise small value (infinite norm sense) can make new image vector further away from the original one in the euclidean space ($l^2$ norm sense). Denoising also generates very different image vector $D(v)$ from the original $v$. However, classifier is rather robust to denoising than adversarial perturbations. This result implies that deep neural models are vulnerable to noise in certain directions, which can be also shown from the experimental results of UAP [6]. The natural question then arises, is there a certain direction that can make classifier perform better? We assume that this 'defensive' perturbation $v_{def}$ can be generated by averaging random perturbations that lead classifier to change their prediction correctly when it was classified incorrectly or to predict more confidently when it was classified correctly as in Equation 1. In the equation, $L_y(v)$ is a loss function for input v and target class y. In the experiment, we use the random perturbations that satisfies the first case.

$$v_{def} = \epsilon \frac{\sum_i \mathbf{I}(L_{y_i}(v_i + \delta_i) < L_{y_i}(v_i))\delta_i}{\sum_i \mathbf{I}(L_{y_i}(v_i + \delta_i) < L_{y_i}(v_i))} \quad \text{where} \quad \delta_i \sim N(0, I) \tag{1}$$


*KDD'18 Deep Learning Day, August 2018, London, UK*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

### 2.1 Denoising methods: ROF and Non-local means

There are several techniques for removing noise from an image. The classical ROF model, also known as 'Total variation denoising', is an algorithm that can smooth the noise by reducing total variation [7]. This noise removal method is widely used because it is simple and fast. However, important features such as texture can often be overly smoothed. This drawback can be resolved by a denoising method called non-local means method. The non-local means algorithm can preserve fine structure since it implements non-local means as in Equation 2 instead of local means. In the equation, $\Omega$ is the image area, $v$ is an unfiltered image, $u$ is the filtered image, $f$ is a weighting function, and $Z(p)$ is the normalization factor:

$$u(p) = \frac{1}{Z(p)} \int_\Omega v(q) f(p,q) dq, \quad \text{where} \quad Z(p) = \int_\Omega f(p,q) dq \tag{2}$$

## 3 EXPERIMENTS

We use 10K ImageNet training data [8] to train the UAP noise, $\delta_{UAP}$ and 10K ImageNet validation data to test the proposed methods. For denoising parameters, we use a weight of 10 for ROF, a patch size of 2, a patch distance of 3, and a cut-off distance of 0.05 for NL-means. As a classifier, we use the pre-trained inception V1 [10] model as target model and use VGG-16 and VGG-19 [9] as black box attack models. A typical example for ROF model is shown in Figure 1.

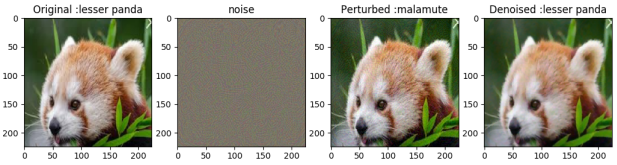


**Figure 1: The ROF model can remove the universal adversarial noise and shows correct classification for the most cases (see details in Appendix). The classified label on the top of each image (True = lesser panda). From left to right: Original image, noise (UAP), Perturbed image, Perturbed image after denoised**

The ROF model minimizes the total variance $J(u)$ remaining the variance $\sigma(u)$ from $v$ is smaller than given constant $\sigma$, so $u$ is in the $l^2$ ball with radius $\sigma$ centered at $v$. Therefore, the result image is heavily dependent on the value of $\sigma$ that we control by denoising weight. We applied different weights (5, 10, 20) and the results are shown in Figure 2, 3, and 4. For NL-means, we control the degree of denoising by changing cut-off distance $h$, which can be considered as a width parameter in Gaussian kernel. Using the NL-means with larger $h$, we get more blurry images. However, the results shown in A implies that the performance is robust to the denoising parameter within the range of concern. For the experiments on the additional defensive noise $v_{def}$, the Equation 1 and 10K ImageNet validation data that is disjoint from the test set are used.

**Table 1: Classification accuracy for the case of clean, perturbed, denoised (ROF, nlm) image, denoised image with additional noise. The target model is IncV1 and each column represents the attack model.**

| | IncV1 | VGG16 | VGG19 |
|---|---|---|---|
| $v_0$ | | 69.4 | |
| $v_0 + \delta_{UAP}$ | 12.5 | 28.7 | 31.7 |
| $D_{ROF}(v_0 + \delta_{UAP})$ | 56.3 | 47.2 | 49.1 |
| $D_{ROF}(v_0 + \delta_{UAP} + v_{def})$ | 57.1 | **51.2** | **51.3** |
| $D_{nlm}(v_0 + \delta_{UAP})$ | 57.7 | 45.0 | 46.9 |
| $D_{nlm}(v_0 + \delta_{UAP} + v_{def})$ | **58.0** | 47.7 | 49.4 |

The results of classification accuracy are shown in the Table 1. The classification accuracy of the Inception V1 target model is 69.4% for the undamaged images. The UAP perturbation trained by Inception V1 model can perform white-box adversarial attack and degrades the accuracy to 12.5%. The VGG-16 and VGG-19 model can train black-box UAP that can degrade the accuracy to 28.7% and 31.7% respectively. The denoising $D_{ROF}$ and $D_{nlm}$ highly increase the accuracy around 45 ~ 58%. The additional noise also improves the classification accuracy in all cases. This result implies that the defensive noise is well trained to reduce the loss and move the input images to high-density area.

## 4 CONCLUSIONS

In this paper, we propose defensive denoising methods against adversarial attack using ROF and NL-means model. We also construct a perturbation that can help the classifier predict better by averaging noises that reduce the loss. In the experiments, our schemes can improve the classification accuracy and the results are rather robust to the denoising parameters.

## 5 ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2016R1A2B3014030) and (MSIP)(No.20 17R1A5A1015626)

## REFERENCES

[1] A. Buades, B. Coll, and J.M. Morel. 2005. A non-local algorithm for image denoising. *Computer Vision and Pattern Recognition(CVPR)* 2 (2005), 60–65.
[2] I. J. Goodfellow, J. Shlens, and C. Szegedy. 2015. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)* (2015).
[3] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)* (2016).
[4] A. Kurakin, I. J. Goodfellow, and S. Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* (2016).
[5] S.M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. 2016. DeepFool: a simple and accurate method to fool deep neural networks. *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)* (2016), 2574–2582.
[6] S.M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. 2016. Universal adversarial perturbations. *arXiv preprint arXiv:1610.08401* (2016).
[7] L. I. Rudin, S. Osher, and E. Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena* (1992).
[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* (2015).
[9] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations(ICLR)* (2014).

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
[11] C. Szegedy, W. Zaremba, I. Sutskever, and J. Bruna. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

# A DENOISING RESULTS

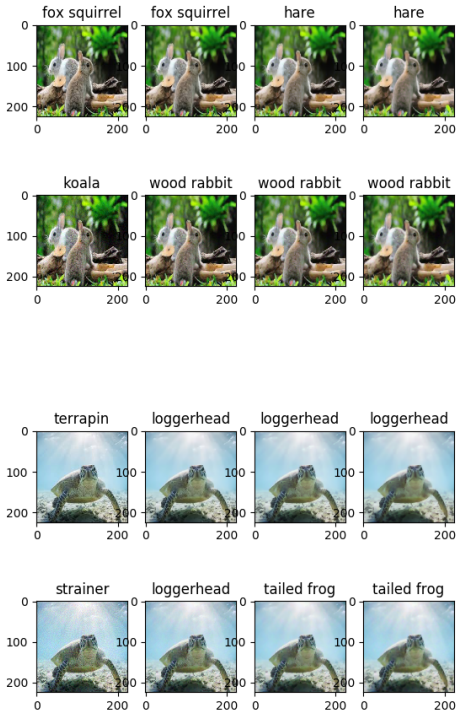


**Figure 2: The weight parameter can change the classification performance. In the most cases, classifier is robust to the weights within the range we consider(see 3 and 4). The classified label on the top of each image. Top: Original, Bottom: UAP perturbed image, From left to right: baseline image, $w = 5$, $w = 10$, and $w = 20$**

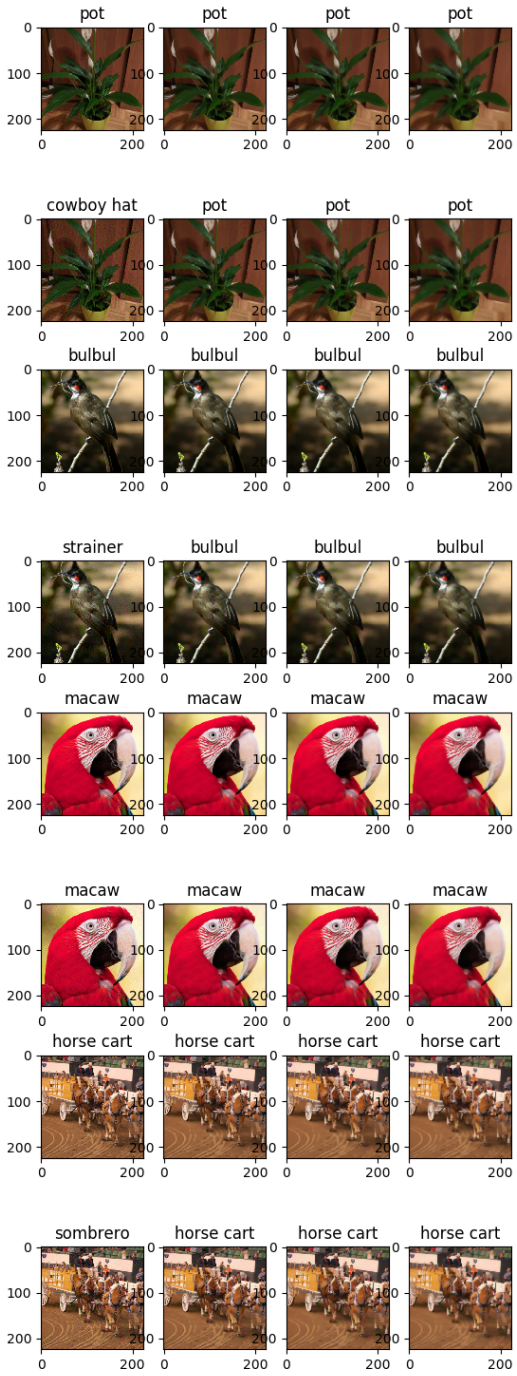


**Figure 3: The experiment result for ROF model with the classified label on the top of each image. Top: Original, Bottom: UAP perturbed image, From left to right: baseline image, $w = 5$, $w = 10$, and $w = 20$**

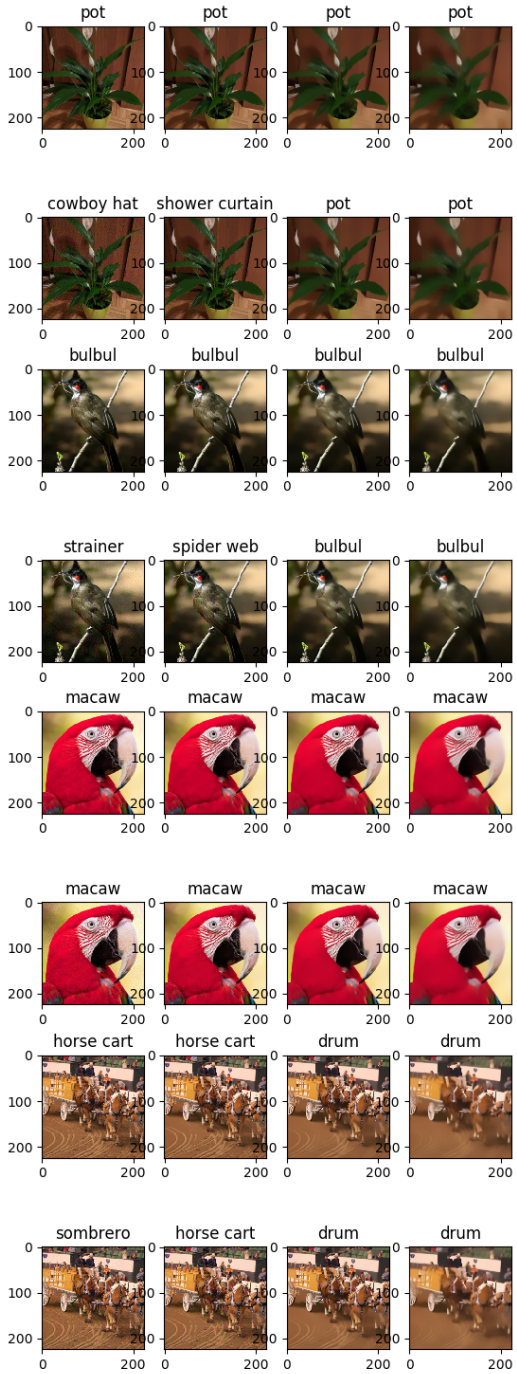


**Figure 4: The experiment result for NL-means model with the classified label on the top of each image. Top: Original, Bottom: UAP perturbed image, From left to right: baseline image, $h = 0.02$, $h = 0.05$, and $h = 0.1$**