Information Redundancy Minimization for Adversarial Defense

XU Xiao[1†], YANG Xingyi [2†], CHEN Yijun[2†], WANG Zheng[1,3], HU Yining[1,3*] and XIE Lizhe[4,5*]

Abstract — Adversarial attacks are becoming serious threat to information security. In this paper, we proposed Information Redundancy Minimization(IRM) method to counter the adversarial attacks. IRM contains two stages: compression and image information minimization, which introduces a multi-scale ensemble model. The advantages of IRM are manifested in three ways: 1) reduced training time. 2)combining multi-scale input with compression methods. 3) compatible with other defense methods. Finally, IRM methods achieves 92.5% accuracy in the IJCAI–19 AAAC datasets, which far higher than 83.8% accuracy achieved by adversarial training defense and 85.9% accuracy achieved by compression defense methods.

This work is supported by Open Project from Jiangsu Key Laboratory of Oral Diseases, Nanjing Medical University (JSKLOD-KF-1701, JSKLOD-KF-1708);

[1]School of Cyber Science and Engineering, Southeast University, Nanjing, 210000, China; [2]School of Computer Science and Engineering, Southeast University, Nanjing, 210000, China; [3]Key Laboratory of Computer Network Technology of Jiangsu Province; [4]Institute of Stomatology,Nanjing Medical University,Nanjing,China; [5]Jiangsu Key Laboratory of Oral Diseases, Nanjing Medical University,Nanjing,China

[†]The first three authors contributes equally to this paper.

[*]Corresponding Author

***Keywords* — adversarial defense, neural network security, information redundancy minimization**

I. INTRODUCTION

Recent breakthroughs in computer vision have brought some new security challenges[1]. The Convolutional Neural Network(CNN) structure has evolved from AlexNet, VGG-net, Inception to Resnet and its variants[2], They have been widely used in security applications such as autonomous vehicles, face recognition and malware detection due to their significant capability of multi-scale feature extraction and generalization.

However, the CNN-based systems can be vulnerable and lucrative targets for attackers. Neural networks have been found vulnerable to subtle input perturbations which lead to completely preposterous outputs[3]. Therefore, making CNN more robust to adversarial examples is a very important but challenging issue.

In general, defenses against adversarial attacks mainly focus on three aspects: training process, network architecture and pre-processing of input examples.

The most straightforward approach is adversarial training[4-5] which directly uses adversarial examples to augment the training set. But these methods require

enormous computing resources and training time. For network architecture, Deep Contractive Network[6] is a generalization of the contractive auto-encoder, which imposes a layer-wise contractive penalty in a feed-forward neural network. Dense Associative Memory model[7] tries to enforce higher order interactions between neurons by changing rectified linear units (ReLU) to rectified polynomials.

Pre-processing is the most migratory methods. And image denoising is one of the most widely used methods. Guo C et al.[8] proposed input transformation to denoise and counter with adversarial attack. And F. Liao et al.[9]assumed defense adversarial samples to be a denoising operation. In addition, reducing image size is also an effective way to defense against adversarial attacks in practice.

Inspired by the ideas of image denoising, we define the features that has little influence on the model's classification accuracy and is easy to hide additive perturbations as redundant information. The process of adversarial attack could be deemed as adding redundant signal to images regardless of the content of itself. The misleading redundant information is usually imperceptible to the human, but cause excessive attention from CNN and cause them to misclassify the manipulated instances with high confidence[10]. Although perturbations are required to be as small as possible, once they are added to several key locations, it may result in unpredictable perturbation at the feature level. Whether the attack model is known, the philosophy of redundancy addition is what they have in common, which means we can defense by suppress such information.

II. THE PROPOSED METHOD

In this paper, we propose Information Redundancy Minimization(IRM) as the defense for adversarial attacks. The overall pipeline is shown in Fig.1. Our methodology involves two stages: first we use JPEG compression to eliminate redundant high frequency component and denoise; then the scheme of down-sampling and spatial pyramid pooling (SPP)[11] is applied to combine multi-scale prediction results, which preserves the performance on non-adversarial images but also improves the defense model's robustness.

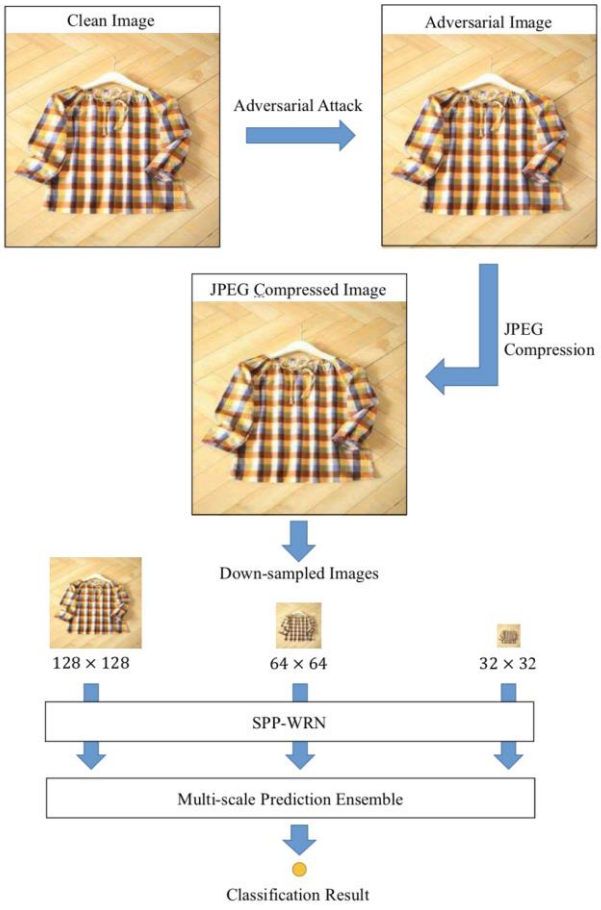

Figure 1 Overall pipeline

A. Random Compression and Merging

Compression is a common denoising method. And JEPG compression has been proved to be an effective way to reduce the high frequency signal of pictures and eliminate the image information that are difficult to perceive by human eyes[12], we use JEPG compression as a method to destroy the additional disturbance caused by the attack and improve the defense performance of

the model.

According to the facts mentioned above, we refer to the scheme of Nilaksh et al.[12], and propose two steps of JPEG compression defense method, as follows:

1. Preprocess the input image by JPEG compression to weaken the additional disturbance generated by the attack.

2. Reconstruct the image by random compression and merging to enhance the generalization ability of the defense model. We choose four different levels of image compression, retaining 40%, 60%, 80% and 90% of the image information respectively. Then the picture is divided into four parts. Each part randomly selects one compression degree for JPEG compression. Finally, the four parts are merged into one complete image. This strategy is shown in Fig.2.

The random compressed patches reduce the high frequency component from different level, therefore provide the defense model higher resistance to adversarial attacks compared to conventional single level JPEG compression..

B. Multi-scale Ensemble with Spatial Pyramid Pooling

Typical classification neural networks call for fixed size image as its input. However, such priori settings can easily be used by attackers to add additional disturbance at a certain scale. Thus, we propose to integrate the multi-scale prediction to minimize the total image information as defense. Firstly, we will verify the impact of different scales on clean images and the impact on adversarial images, and then select the appropriate scale for multi-scale integrated prediction

Scale Variance under Clean Images

First, we analyze the CNN's performance on clean samples at different scales. In the experiment, we trained three Wide Residual Networks(WRN) models with input

Figure 2 Compression strategy

size of 32, 64, and 128 on the IJCAI-2019 dataset through Decoupling Direction and Norm (DDN)[13] as attack method, and tested the classification accuracy of clean samples at different scales. In the Tab.1, column represents a model trained at a certain scale and row represents the size of the input image at the time of the test. By fixing other factors, we may observe that: (1) The model trained at a certain scale will have a large precision loss when testing on images of other scales. (2) The greater the training resolution, the higher accuracy on clean images can be reached.

Scale Variance under Adversarial Images

Then we analyze the DNN's performance on adversarial samples at different scales. We use the same models on clean image and use Fast Gradient Sign Method(FGSM)[10] to attack the test-set on (299 × 299) The classification accuracy shown in Tab.2 indicate that training with high resolution images results in sensitive classifiers to FGSM attacks and lower classification accuracy on adversarial samples.

Multi-scale Ensemble for Defense

According to the experiments above, we may come to the conclusion: For the same network structure, the model trained with higher resolution tends to gain higher classification accuracy on the clean samples, yet with poor performance on the adversarial images. Such experimental results are consistent with human visual cognition habits: in general, high resolution images are easier to classify, while low resolution images are blurred and indistinguishable. However, due to the existence of certain information redundancy in large-scale images, it is more sensitive to adversarial attacks; on the contrary, reducing the size or resolution of the input image can improve the robustness of the model. In compromise between accuracy and robustness of the classification network for both clean and adversarial samples, we use spatial pyramids pooling (SPP) to perform multi-scale prediction and integration.

Table 1 Classification accuracy of clean samples at different scales

Train / Test	32*32	64*64	128*128
32*32	90%	69.09%	24.54%
64*64	36.36%	92.27%	75.45%
128*128	27.63%	59.45%	95.32%

Table 2 Classification accuracy of adversarial samples at different scales

Train / Test	32*32	64*64	128*128
32*32	89%	63.09%	17.64%
64*64	26.43%	82.27%	68.35%
128*128	21.63%	48.45%	75.43%

We replace the last pooling layer with a spatial pyramid pooling layer in order to enable multi-scale input to the network. The input feature map of size $M \times N$ is divided into $k_i \times k_i$ bins, where $k_i \in \{1, 2, 4\}$ in our method. Each bin is pooled (in this paper we use average pooling) with a filter of its own size. The fixed-dimensional output vectors are fed to the fully-connected layer for classification. With spatial pyramid pooling, we can resize the input images to any scale and apply the same deep network. Fig.3 shows the diagram of spatial pyramid pooling.

In the training process with SPP layer, we use a method similar to multi-scale training, resizing each image to three scales and adopt back-propagation for each input. During the test step, three down-sampled images are predicted separately. We integrated the prediction results by weighted summation of confidence, visualized in Fig.4. As equation (1) goes,

$$f(x) = \sum_{i=\{32,64,128\}} \lambda_i f(x_i) \tag{1}$$

where $f(x)$ is the prediction result, λ_i refer to the scale importance weights and x_i is to resized image of $i \times i$, in which $i \in \{32, 64, 128\}$.

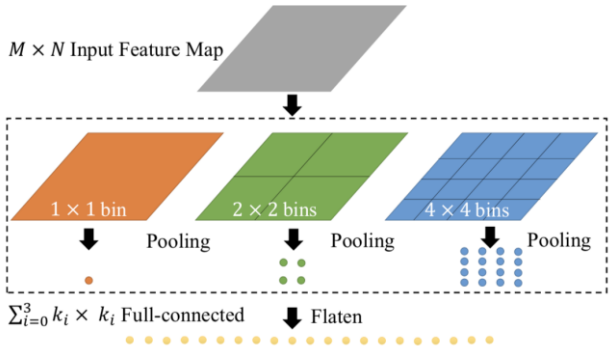

Figure 3 The diagram of spatial pyramid pooling

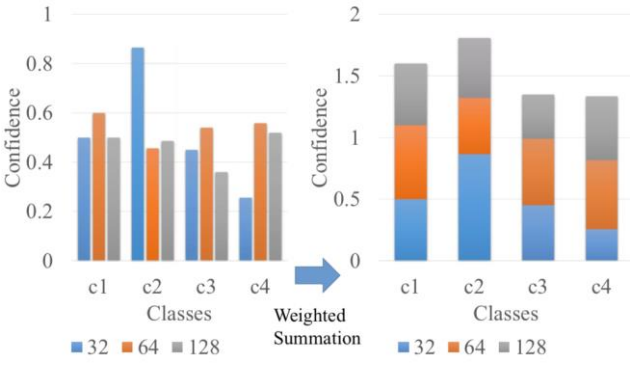

Figure 4 Multi-scale prediction ensemble with weighted

summation of confidence. The classification confidence under each scale is added with weights to get the final prediction

In our approach, the scale weights is a critical issue. We use grid search for select these hyper parameters. The search range is {0.001, 0.01, 0.1, 1, 10}. The experimental results show that when $\lambda_{32} = 1$, $\lambda_{64} = 0.01$, $\lambda_{128} = 0.001$, the attack can be most effectively defended.

Unlike the image pyramid, our prediction process uses only a single model, but each image needs to be down-sampled to three scales for three predictions. Such mechanism can improve the adaptability of the network against the adversarial samples without performance loss on clean images.

Another advantage of IRM is that the network does not need to train from scratch. For the pre-trained classification model, we only need to replace the pooling and the full-connected layer to continue training, without necessity to alter the structure of feature extractor.

C. *Implementation Framework*

The training process is composed of 3 steps.

1. Adversarial training is considered, which considers augmenting the training objective with adversarial examples[14], with the intention of improving robustness. Suppose the input clean image is x and the label is y. Given a model with loss function $J(x, y, \theta)$, the training is augmented as (2):

$$\tilde{J}(x, y, \theta) = \alpha J(x, y, \theta) + (1 - \alpha) J(\tilde{x}, y, \theta) \quad (2)$$

where $\tilde{x}$ is an adversarial sample and α is the weight. The model structure is shown in Fig.5. DDN[13] is the basic attack for our methods to generate adversarial samples. Under a certain number of iterations, the DDN model searches for the optimal disturbance mask and normalizes it, whose objective is to obtain the worst possible loss for a given maximum noise of norm ϵ, the optimization procedure to obtain an attack with minimum distortion δ can be formulated as (3):

$$min_{\delta} \ P(y_{true}|x + \delta, \theta)\}$$
$$\delta \leq \epsilon \ and \ 0 \leq x + \delta \leq M \quad (3)$$

Where $P(y_{true}|x + \delta, \theta)\}$ is the accuracy of classification model and M is the range of image pixel value. DDN enables a novel adversarial training. At each iteration, we train with examples close to the decision boundary.

2. Based on the adversarial training model, we apply IRM to start a new training round. In order to separately verify the validity of each part of our approach, we implement our method as follows. We train the multi-scale ensemble CNN model with SPP layer, and then combine with JPEG compression to get the final model. The model structure is shown in Fig.6
3. We improve the stability of the model by modifying loss and sample integration.

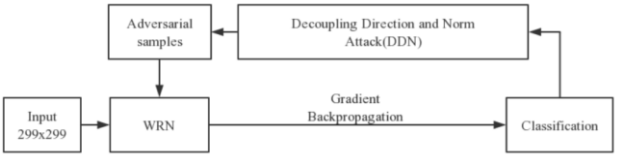

Figure 5 Adversarial training model based on wide residual net-work and DDN attacks

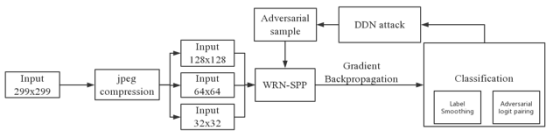

Figure 6 IRM method structure includes JPEG compression and SPP

Label Smoothing: Label smoothing[15] use soft targets for the cross-entropy loss rather than hard targets. The correct class is given a target probability of $1 - \delta$ and the remaining δ probability mass is divided uniformly between the incorrect classes. Because smaller logic usually leads to a smoother output

distribution, and Goodfellow et al.[5] indicated that label smoothing provides a small amount of robustness for adversarial examples.

Adversarial logit pairing: We add a regularization term according to the adversarial logit pairing (ALP)[15] scheme on the basis of cross-entropy function so as to enable the model distinguish clean and adversarial samples. ALP matches the logits from a clean image *x* and its corresponding adversarial image $\tilde{x}$. In traditional adversarial training, the model is trained to assign both *x* and $\tilde{x}$ to the same output class label, but the model does not receive any information indicating that $\tilde{x}$ is more similar to *x* than to another example of the same class. The extra regularization term can encourage similar embeddings of the clean and adversarial versions of the same example, helping to guide the model towards better internal representations of the data. The final loss function is defined as (4):

$$J(M,\theta)+\alpha\frac{1}{m}\sum_{i=1}^{m}L\left(f\left(x^{(i)};\theta\right),f\left(\tilde{x}^{(i)};\theta\right)\right) \quad (4)$$

where $f\left(x^{(i)};\theta\right)$ is the function mapping from inputs to logits of the model. And $J(M,\theta)$ is the cost function used for adversarial training.

Ensemble Training: By amplifying the attack samples such as Projected Gradient Descent (PGD) [15] and DeepFool(DF)[10], incremental training is performed to enhance the generalization ability of the model. PGD and DF are all untargeted attacks with typical characteristics.

III. EXPERIMENTAL EVALUATION

Dataset and metrics

IJCAI-2019 Alibaba Adversarial AI Challenge (AAAC 2019) datasets releases totally 110,000 online ecommerce images, which come from 110 categories of products. However, due to the time limitation, we only use 50% of datasets as training set and 10% for validation, including 110 categories, sized 299 × 299 × 3 (the provisions of the competition test set) in the following experiments.

The challenge defines a normed score as the online evaluation metrics. For an adversarial image, if our defense model is misled, the score is 0; if the image is classified correctly, the average L2 distance between the adversarial sample and the original sample is calculated as the score. The score $D(x,\tilde{x})$ is defined in (5).

$$D(x,\tilde{x})=\begin{cases}0\ if\ f(\tilde{x})\neq y\\ mean(\|x-\tilde{x}\|_2)\ if\ f(\tilde{x})=y\end{cases} \quad (5)$$

Experimental Setting

The hyper parameters of WRN classification network is 28 in width and 10 in depth with drop rate of 0.3. As shown in Tab.3, network width is determined by factor *k=10*. Groups of convolutions are shown in brackets where *N=4* is a number of blocks in group, down-sampling performed by the first layers in groups conv3 and conv4. Final classification layer and average pooling layer are omitted for clearance. The validation set during training only contains clean images. The initial learning rate is 0.01 but it decreases to 0.0001 as the training epochs increase. The optimization method is Stochastic Gradient Descent (SGD). Bach size is set to 16 according to the computing resource. The basic classifier's accuracy on clean images can reach up to 95.32%. Then, we used DDN to retrain the model with the weight of WRN classification model, which can speed up the rate of converge. Taking into account the limited computing resource, the number of DDN's attack steps is 5 while the disturbance is below 1 basically.

In our experiment, FGSM attack using Foolbox[16] is chosen as a black box attacker to generate adversarial images as evaluation datasets to evaluate defense models' performance, which can prove our model's transferability.

Our experiment runs on ubuntu operating system with GTX 1080ti GPU and 16G memory.

Table 3 Structure of 28-10 wide residual networks (k =10, N=4)

group name	output size	block type=B(3,3)
conv1	32 × 32	[3 × 3 , 16]
conv2	32 × 32	$\begin{bmatrix} 3\times 3, & 16\times k \\ 3\times 3, & 16\times k \end{bmatrix}\times N$
conv3	16 × 16	$\begin{bmatrix} 3\times 3, & 32\times k \\ 3\times 3, & 32\times k \end{bmatrix}\times N$
conv4	8 ×8	$\begin{bmatrix} 3\times 3, & 64\times k \\ 3\times 3, & 64\times k \end{bmatrix}\times N$

Evaluation Result

The fundamental of IRM is that image information decreasing removes the adversarial effect. Fig.7 compares the classification accuracy on adversarial samples between WRN-SPP-DDN model under multi-scale training and WRN-DDN baselines. We may notice that introducing SPP in the multi-scale training may cause inferior performance for each single scale. But in combined analysis with Tab.4, we can see that the multi-scale prediction integration on FGSM attacks can achieve an accuracy of 91.6%. Generally, the smaller image scale can guarantee the performance against attacks. These figures show that multi-scale prediction integration can effectively improve the adaptability of the model to the adversarial samples, and improve the overall robustness of the model.

Tab.4 compares the evaluation results of various methods. The results demonstrate that the proposed IRM method can significantly improve the accuracy based on WRN-DDN. The multi-scale prediction integration with SPP significantly promotes the accuracy from 83.8% to 91.6%; the JPEG compression is also positive to adversarial defense, with an accuracy elevation of 2.1%. Combination of jpeg compression and multi-scale SPP method further improves the robustness of the model, the accuracy reach 92.5%, which is much higher than the baseline performance and general compression methods.

Additionally, we implement the commonly used ensemble adversarial training, ALP and other techniques on top of our method, and the adaptability of the model can be further enhanced, reaching the final competition score of 19.2, which means that the method in this paper is compatible with common defense techniques.

It is worth notice that, our final approach can also reach 94.8% on clean test images. It means that our proposed method can ensure robustness against strong attacks as well as to maintain satisfying performance on clean samples.

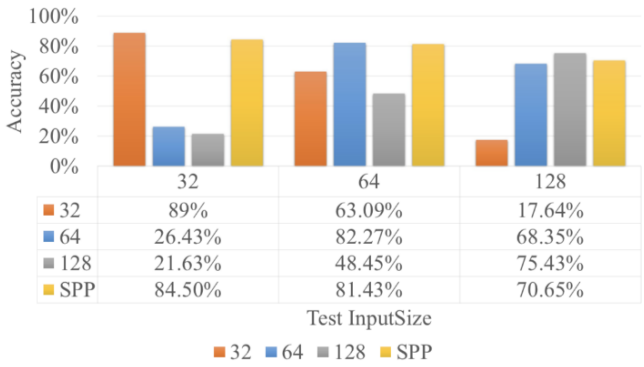

Figure 7 Accuracy of Adversarial image With SPP. Orange, Blue and Gray bars are baseline models trained on 32×32, 64×64, 128× 128 images respectively, yellow bar is trained on multi-scale images with SPP-layer

IV. CONCLUSION

In this paper, we propose a method called IRM to mitigate adversarial effects. On one hand, we conducted comprehensive experiments to find proper scale spatial pyramid pooling (SPP) of various scales so as to minimize the redundancy. On the other hand, we use traditional denoising methods such as JPEG compression to reduce the high frequency components, which contributes to reduce training time and improve the defense performance.

The main contributions of the proposed work is: we propose IRM as a novel defense method on the basis of adversarial training, which combines the random compression and spatial pyramid pooling to integrate

multi-scale prediction. And results illustrate the effectiveness of IRM to eliminate additional disturbances of attacks. The evaluation result demonstrates that IRM combined with adversarial training can reach 92.5% on adversarial test samples, while 83.8% for adversarial training baseline.

The advantages of IRM are as follows:

- Combining multi-scale input with compression method， better defensive effect against adversarial attack is achieved.
- Compared to conventional integration on multi-scale detections, the proposed IRM requires fewer computation time. The average processing time for one sample sized 299 × 299 × 3 is about 14ms, while that of the integration on multi-scale detections is over 30ms.
- Our method is compatible to adversarial defense methods, which can serve as an additional module for adversarial defense.

The method can also be integrated with multiple models with random weights to further enhance the robustness of the defense model. The experimental results on public dataset show that the minimization of image redundancy information is effective.

Table Classification performance comparison by accuracy and normalized score on IJCAI-AAAC 2019 Datasets

Baseline Model		IRM method		Other Tricks	Accuracy	Score
WRN	DDN	Compression	Multi-scale SPP			
√					60.3%	7.9
√	√				**83.8%**	**15.8**
√	√	√			**85.9%**	**16.2**
√	√		√		91.6%	18.4
√	√	√	√		**92.5%**	**18.5**
√	√	√	√	√	94.0%	19.2

References

[1] 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.

[2] Zagoruyko and N. Komodakis, “Wide residual networks,” in British Machine Vision Conference 2016, British Machine Vision Association, 2016.

[3] N. Akhtar and A. Mian, “Threat of adversarial attacks on deep learning in computer vision: A survey,” IEEE Access, vol. 6, pp. 14410–14430, 2018.

[4] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” arXiv preprint arXiv:1611.01236, 2016.

[5] F. Tram`er, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” arXiv preprint arXiv:1705.07204, 2017.

[6] S. Gu and L. Rigazio, “Towards deep neural network architectures robust to adversarial examples,” arXiv preprint arXiv:1412.5068, 2014.

[7] D. Krotov and J. Hopfield, “Dense associative memory is robust to adversarial inputs,” Neural computation, vol. 30, no. 12, pp. 3151–3167, 2018.

[8] Guo C, Rana M, Cisse M, et al. Countering Adversarial Images using Input Transformations[J]. 2017.

[9] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, “Defense against adversarial attacks using highlevel representation guided denoiser,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1787, 2018.

[10] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, “Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression,” arXiv preprint arXiv:1705.02900, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

[12] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. E. Kounavis, and D. H. Chau, "Shield: Fast, practical defense and vaccination for deep learning using jpeg compression," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 196–204, ACM, 2018.

[13] J. Rony, L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger, "Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses," arXiv preprint arXiv:1811.09600, 2018.

[14] 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.

[15] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," arXiv preprint arXiv:1803.06373, 2018.

[16] J. Rauber, W. Brendel, and M. Bethge, "Foolbox: A python toolbox to benchmark the robustness of machine learning models," arXiv preprint arXiv:1707.04131, 2017.

About the authors

XU Xiao was born in Xuzhou, China. He received the bachelor degree in information engineering from Southeast University, China. He is now a Master and specialize in cyberspace security in Southeast University. His research interests include computer vision and machine learning, especially classification and neural network Security

YANG Xingyi was born in Chongqing, China. He received the bachelor degree in computer engineering from Southeast University, China. He is now a master student of electrical and electronics in University of California, San Diego. His research interests include computer vision and machine learning theory

CHEN Yijun was born in Jiangxi Province. She received the bachelor's degree in the major of Internet of things from University of Science and Technology Beijing. She is now a Master and specialize in computer science and technology in Southeast University. Her research interests include computer vision and machine learning, especially classification and objection detection.

WANG Zheng received his Ph.D in Bio-engineering from Southeast University, China in 2006. He is currently a senior lecturer in School of Cyber Science and Engineering, Southeast University, China. His research interest is image processing, micro imaging and machine vision.

HU Yining received his Ph.D in Bio-engineering from Southeast University, China in 2009. He is currently an associate professor in School of Cyber Science and Engineering, Southeast University, China. His research interest is computer vision, machine learning and cyber security (Email: hyn.list@seu.edu.cn)

XIE Lizhe received his Ph.D in Bio-engineering from Southeast University, China in 2012. She is currently a senior lecturer in Institute of Stomatology, Nanjing Medical University, Nanjing, China. Her research interest is computer vision in medical imaging and medical image analysis. (Email: xielizhe@njmu.edu.cn）