Improving Robustness of Medical Image Diagnosis with Denoising Convolutional Neural Networks

Fei-Fei Xue[1,2⋆], Jin Peng[1⋆], Ruixuan Wang[1,2], Qiong Zhang[1,3], and Wei-Shi Zheng[1,2]

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China
[3] Guangdong Key Laboratory of Information Security Technology

Abstract. Convolutional neural networks (CNNs) are vulnerable to adversarial noises, which may result in potentially disastrous consequences in safety or security sensitive systems. This paper proposes a novel mechanism to improve the robustness of medical image classification systems by bringing denoising ability to CNN classifiers with a naturally embedded auto-encoder and high-level feature invariance to general noises. This novel denoising mechanism can be adapted to many model architectures, and therefore can be easily combined with existing models and denoising mechanisms to further improve robustness of CNN classifiers. This proposed method has been confirmed by comprehensive evaluations with two medical image classification tasks.

Keywords: Robustness of CNN, Adversarial Noises, Denoising CNN, Skin Disease, Chest X-ray.

1 Introduction

Convolutional neural networks (CNN) have been widely used in medical image analysis, such as automatic segmentation of tumor regions in MRI [13] and intelligent diagnosis of skin cancers [4]. However, the application of a medical analysis system would be limited if it is sensitive to various noises and varying environments. One way to critically evaluate the robustness of a medical image system is by adversarial attacks [14]. Specifically, clean images can be altered with imperceptible perturbations (called adversarial noises) to generate adversarial examples, and such adversarial examples can fool CNN classifiers to make incorrect predictions with high confidence. Recent studies on natural images clearly demonstrate that CNN classifiers can be easily attacked and become completely crashed [8]. Adversarial attacks have also been performed on medical images [12], confirming the high sensitivity of CNN diagnosis systems to

⋆ The authors contribute equally to this paper.

adversarial noises. Therefore, it is highly demanding to improve the robustness of intelligent diagnosis systems.

System robustness can be improved by providing the system with the ability of defending adversarial attacks. Multiple defense approaches have been proposed for this purpose. For example, adversarial training and its variants can improve system's defense ability simply by adding one or more types of adversarial examples into the training data during classifier training [5,8,15], while denoising approach pre-processes images often with certain type of auto-encoders, aiming to remove potential adversarial noises before inputting images to classifiers [1,10]. Adversarial training requires embedding adversarial attacking process into classifier training [5,8,15], and denoising approach often suffers from accuracy reduction in classifying clean images [1,10]. Another approach is to train a distillation network which can improve defense ability by effectively enlarging gaps between distributions of classes in the high-level semantic feature space [11].

While the attacking and defense studies of deep neural networks have been actively investigated on natural images in the past several years, few work has investigated the robustness of medical image analysis associated with its defense ability. This paper proposes a novel defense strategy to improve the robustness of intelligent diagnosis systems. Different from existing approaches, the proposed method directly improves network classifier's denoising ability with a naturally embedded auto-encoder and a semantic feature invariance strategy for general noises. This novel denoising mechanism can be adapted to many classifier architectures and is independent of any image pre-processing procedure. Therefore, it can be easily combined into the existing models and denoising mechanisms to further improve the robustness of network classifiers. Experiments on a skin image dataset and a chest X-ray dataset demonstrate that, it can always significantly improve the robustness of the classifiers via integrating the proposed denoising mechanism into the existing CNN classifiers, no matter whether the classifiers have employed other defense methods.

2 Methods

In adversarial attacks, image pixel values can be manipulated via small and carefully-crafted perturbations, such that the originally imperceptible adversarial noise can be progressively amplified over layers in deep neural networks, leading to incorrect classifications. Consider an image as a point in the original high-dimensional image space, and the re-ordered output of the last convolutional layer in a CNN classifier as a point in a low-dimensional high-level semantic feature space. For an original (clean) image which can be correctly classified by the neural network classifier, the corresponding adversarial example should be in a small hypersphere centered at the clean image in the image space, while the two images should be relatively far from each other in the high-level semantic feature space due to mis-classification of the adversarial example. To defend attacks from adversarial examples, one intuitive idea is to assure that the convolutional layers in the classifier can transform all neighboring points around each

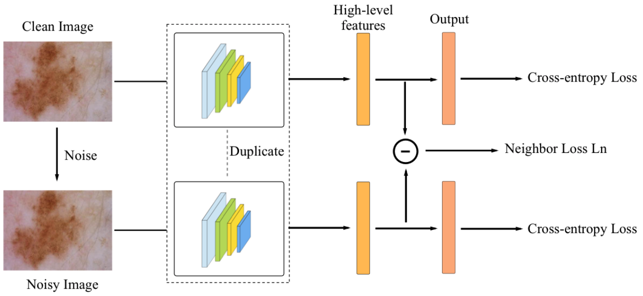

Fig. 1. Training a CNN classifier with an additional loss to emphasize similarity between noisy images and clean images in the semantic feature space. Each noisy image is generated by adding random noise to the corresponding clean image.

clean image to the same point in the semantic feature space as that of the clean image. The popular adversarial training, which adds adversarial examples to the training set, can be considered as one simplified implementation of this idea. Another idea is to remove (both general and adversarial) noises by projecting images to and then conducting reconstruction from a lower dimensional space, supposing that the clean images lie on a low-dimensional manifold while the noisy images are not. With this idea, auto-encoder has been applied to process images often before they are fed into the classifier. By combining the two ideas, but without using adversarial examples and the pre-processing procedure, we propose a novel plug-and-play mechanism to defend adversarial attacks, thus improving system robustness.

2.1 Transforming neighboring noisy images to the same point

Because each adversarial example falls within a small neighborhood of the corresponding clean image in the image space, the classifier trained additionally with all the available noisy images, within the neighborhood of each training image, should become more robust to adversarial attacks, in the sense that all noisy images around a clean image would be projected to the same point in the semantic feature space. However in practice, it is infeasible to collect all such noisy images. Here by trying to project a small subset of general noisy images within the neighborhood of each clean image to the same point as that of the clean image in the semantic feature space, we expect adversarial examples within the neighborhood would be more likely projected to the same point as well. In this case, the adversarial examples would be more likely recognized as the same class of the clean image, thus improving robustness of the classifier.

Formally, for the i-th clean image $\boldsymbol{x}_i$ in the original training set, let us denote by $\boldsymbol{x}'_i$ a noisy image generated by adding uniform random noise $[-\sigma, \sigma]$ to each pixel of the clean image $\boldsymbol{x}_i$, and $\boldsymbol{f}(\boldsymbol{x}_i)$ and $\boldsymbol{f}(\boldsymbol{x}'_i)$ be the corresponding semantic feature vectors generated by the output of the final convolutional layer in the classifier. Then the objective of transforming neighboring noisy images to the same point in the semantic feature space can be formulated as an optimization problem, i.e., training the classifier (see Fig. 1) such that the following loss function L_n (called neighbor loss) is minimized:

$$L_n = \frac{1}{N} \sum_{i=1}^{N} \| \boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}(\boldsymbol{x}'_i) \|. \tag{1}$$

Note that $\boldsymbol{x}'_i$ can be randomly generated over training iterations such that multiple different noisy images are used for each clean image. Using random noise rather than adversarial noise during model training is one key difference between our approach and existing ones (e.g., [9]).

2.2 Embedded auto-encoder

The existing denoising approaches often employ a separate auto-encoder to remove potential adversarial noise from images before sending image data to the classifier. However, fine details in normal regions in the images could also be modified by the auto-encoder. Such change in normal regions actually causes new noises compared to the original clean images, and these new noises may be progressively amplified over layers in the classifier. Just as the adversarial noises, such new noises may also lead to mis-classification of the images, which actually has been observed in related studies [10] and in our experiments.

To avoid the downgraded classification performance on clean images and meanwhile make use of denoising ability from auto-encoder, we propose to embed the auto-encoder into the network classifier, where the encoder part shares low-to-middle convolutional layers of the classifier (Figure 2 Left). Because the auto-encoder denoises images mainly by projecting them to a lower-dimensional space via the encoder part, sharing the encoder with the CNN classifier would naturally transfer the denoising competence to the classifier. Meanwhile, the classifier still uses original images rather than reconstructed images from the auto-encoder as input. This is clearly different from the existing approach which used a separate auto-encoder before the CNN classifier [9] (Figure 2 Right).

Thus, for the clean image $\boldsymbol{x}_i$ and one corresponding noisy image $\boldsymbol{x}'_i$, with their reconstructed results $\hat{\boldsymbol{x}}_i$ and $\hat{\boldsymbol{x}}'_i$ from the embedded auto-encoder, the classifier can be trained to not only improve the classification performance, but also help improve the reconstruction performance of the embedded auto-encoder by additionally minimizing the reconstruction error L_a:

$$L_a = \frac{1}{N} \sum_{i=1}^{N} \{ \| \boldsymbol{x}_i - \hat{\boldsymbol{x}}_i \|_2 + \| \boldsymbol{x}_i - \hat{\boldsymbol{x}}'_i \|_2 \}. \tag{2}$$

Note that the target of the reconstructed noisy image $\hat{\boldsymbol{x}}'_i$ is the clean image $\boldsymbol{x}_i$. Similarly as in Equation (1), multiple noisy $\boldsymbol{x}'_i$ can be randomly generated for each clean image.

Combining both ideas (Equations 1 and 2), a more robust classifier can be obtained by simultaneously training the classifier and the embedded auto-encoder,

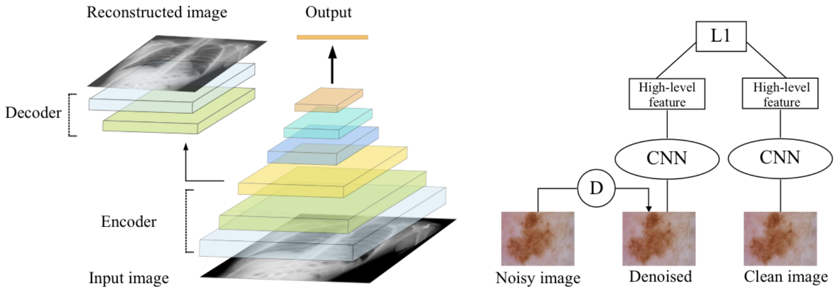

Fig. 2. The proposed embedded auto-encoder with a CNN network (Left) is different from the recently proposed high-level feature guided denoiser [9] (Right). D represents auto-encoder, and L1 represents L_1 distance.

with the constraint to make clean and noisy images similar to each other in the semantic feature space, i.e., by minimizing the loss function L,

$$L = L_c + \lambda_n L_n + \lambda_a L_a \, . \tag{3}$$

Here L_c denotes the cross-entropy loss for the classifier itself to improve its classification performance on both clean and noisy images, λ_n and λ_a are hyper-parameters to respectively control the relative weights of loss terms L_n and L_a.

3 Experiments

3.1 Experimental settings

The experiments were performed on two medical image datasets, the skin image dataset from ISIC2018 Challenge with 7 disease categories(SKIN4) [3], and the chest X-ray dataset with 3 categories [4]. To reduce the data imbalance between classes in the skin dataset, four classes (MEL, NV, BCC and BKL) in which the number of images exceeded 500 were selected and 1,500 images of NV class were randomly selected from 6,705, while keeping the other classes of data unchanged. The selected images were split to training set and test set with the rate around 5:1. For the chest X-ray dataset, we randomly split the raw training set to two parts, with 21,000 images as our training set and the left 6,000 images as test set. Also, to generate adversarial images for the evaluation of the proposed defense approach, different attacking methods, including the Fast Gradient Sign Method (FGSM)[5], the iterative FGSM (IFGSM) [7], and the Carlini&Wagner [2] method (C&W), were applied to four widely used network classifiers, including ResNet18 [6] and VGG-16. For each training sample, we generated two adversarial examples with the perturbation level ϵ in $\{4, 8\}$ for FGSM and IFGSM. And for C&W, we set the searching times to 5 and the iteration times to 1000.

[4] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

There are mainly two types adversarial attacks based on different assumptions on the knowledge of the target network, i.e., white-box and black-box attacks. In the black-box attack, an attacker can observes only the network's output information on some probed input information, which is more realistic and applicable. In comparison, in the white-box attack, an attacker has detailed information on the network architecture and model parameters. The evaluations here mainly focus on the defense of black-box attacks.

All the CNN classifiers used in experiments were optimized using SGD, with initial learning rate set 0.01, and weight decay set 0.0001. Each model was trained on a single GPU with batch size 64. The number of training epochs was set 80. Note that due to limited space, only part of the evaluation results were shown below, and the attacking model was ResNet18 unless otherwise mentioned.

Table 1. Classification accuracy on the adversarial examples generated from the SKIN4 test set. Rows 2 to 4 indicate the influence of neighbor loss and rows 5 to 7 represent the influence of reconstruction loss. Clean stands for original clean images. NA means no defense.

Defense		Clean	FGSM		IFGSM10		C&W
λ_n	λ_a		$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 4$	$\epsilon = 8$	
NA	NA	84.28	19.27	21.99	3.78	3.78	1.54
0.1	-	81.32	41.84	30.97	49.29	30.61	16.31
1		**82.74**	**46.34**	**32.51**	**56.15**	**38.65**	**21.28**
10		64.78	-				
-	0.1	83.22	39.13	27.54	45.27	28.49	15.25
	1	82.62	41.02	29.20	53.43	33.57	19.39
	10	**82.39**	**41.66**	**29.55**	**49.41**	**31.80**	**18.91**
1	**10**	**78.96**	**57.92**	**45.86**	**65.37**	**54.37**	**35.22**

3.2 Evaluations on skin dataset

This section evaluates the effect of the proposed approach in improving robustness of a CNN classifier ResNet18 with ablation study on the skin dataset. Table 1 showed that when including the neighbor loss during training, with the embedded auto-encoded excluded, the trained classifiers (rows 2-4) performed significantly better than the classifier without any defense (first row), when attacked by different methods at different perturbation levels (ϵ). It also shows that with increasing weight λ_n of the neighbor loss, the defense performance increased accordingly. However, large λ_n (10.0) might lead to downgraded performance in classifying clean images. This is reasonable because larger λ_n would make the network pay less attention to the cross-entropy loss during training. As a trade-off, $\lambda_n = 1$ was chosen for subsequent tests on the skin dataset. Note that the decrease in classification accuracy on clean images is a common phenomenon in most defense methods (e.g., see [7,10]).

Similarly, by adding only the embedded auto-encoder to the classifiers, with the neighbor loss excluded during training, Table 1 (fifth row to second last row) showed the trained classifiers also performed significantly better than the classifier without any defense (first row) at various attacking scenarios. As a trade-off, $\lambda_a = 10$ was chosen for subsequent tests. Note that $\lambda_a = 100$ lead to downgraded performance in classifying clean images.

By combing both the neighbor loss and the embedded auto-encoder into the classifier, the trained classifier showed superior performance than all the above results (Table 1, last row), suggesting that the two proposed two ideas work together to further improve the robustness of the classifier. Similar results were obtained on the chest X-ray dataset (Table 2).

Table 2. Classification accuracy of the classifiers on the adversarial examples generated from the chest X-ray test set.

Defense		Clean	FGSM		IFGSM10		C&W
λ_n	λ_a		$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 4$	$\epsilon = 8$	
NA	NA	76.00	20.86	20.91	28.21	16.13	12.39
1	0	75.27	68.48	63.07	71.84	68.68	60.29
0	10	74.85	67.66	62.04	70.66	68.34	57.74
1	**10**	**73.97**	**71.21**	**68.66**	**72.35**	**70.79**	**60.87**

Table 3. Classification accuracy of the architecture on the SKIN4 test set with modified model based on VGG16. The attacking model is ResNet18.

Defense		Clean	FGSM		IFGSM10		C&W
λ_n	λ_a		$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 4$	$\epsilon = 8$	
NA	NA	84.28	38.77	28.25	50.71	29.08	15.60
1	0	82.74	68.68	58.75	74.00	69.50	50.12
0	10	82.39	66.55	56.03	73.05	67.85	44.21
1	**10**	**78.96**	**70.21**	**64.18**	**74.00**	**70.45**	**54.73**

3.3 Combinations with different model structures

To show that our approach can work with different model structures, we combined our idea with another model VGG-16. Table 3 again showed that the proposed defense approach improved the robustness of the CNN classifier with a different structure, compared to the classifier without using defense (row with 'NA'). Combined with the evaluations on the ResNet18 structure above, it supports that the proposed approach helps improve robustness of multiple CNN model structures.

3.4 Combinations with existing defense approaches

To show that our approach is complementary to existing defense approaches, we combined our approach with two existing approaches, the Reformer approach [10] and the HGD approach [9]. Table 4 clearly showed that when one or both of our ideas ($L_n, L_a, L_n + L_a$, corresponding to the neighbor loss, the embedded auto-encoder, or both) were combined with the existing two approaches, the combination further improved the robustness of the classifiers compared to the performance from the existing approaches alone.

Table 4. Classification accuracy when ours combined with existing approaches. The results are based on the SKIN4 test images. R denotes the Reformer approach.

Defense	FGSM		IFGSM10		C&W
	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 4$	$\epsilon = 8$	
NA	19.27	21.99	3.78	3.78	1.54
Reformer(Pixel-level Denoiser [10])					
R	30.16	27.44	18.64	14.41	9.71
R + L_n	48.37	36.85	58.17	42.14	26.06
R + L_a	42.71	31.28	53.47	33.49	25.46
R + $L_n + L_a$	**59.17**	**46.48**	**69.66**	**55.54**	**44.71**
HGD*(High-level features Guided Denoiser [9])					
HGD*	41.80	36.97	42.49	26.01	22.73
HGD* + L_n	56.42	57.01	63.52	46.32	36.77
HGD* + L_a	51.14	47.41	58.18	41.34	31.59
HGD* + $L_n + L_a$	**64.13**	**51.83**	**70.52**	**58.67**	**49.55**

4 Conclusion

In this paper, we proposed a novel defense mechanism to improve robustness of medical image classification systems. This mechanism embeds an auto-encoder into the CNN structure and keeps high-level features invariant to general noises. It is complementary to existing defense approaches and therefore can be combined together to further improve the robustness of CNN classifiers.

Acknowledgement. This work is supported in part by the National Key Research and Development Plan (grant No. 2018YFC1315402) and by the Guangdong Key Research and Development Plan (grant No. 2019B020228001).

References

1. Akhtar, N., Liu, J., Mian, A.: Defense against universal adversarial perturbations. In: CVPR. pp. 3389–3398 (2018)
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy. pp. 39–57 (2017)
3. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N.K., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging, hosted by the international skin imaging collaboration. In: ISBI. pp. 168–172 (2018)
4. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)
5. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)

7. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv:1607.02533 (2016)
8. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial machine learning at scale. CoRR: abs/1611.01236 (2016)
9. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: CVPR. pp. 1778–1787 (2018)
10. Meng, D., Chen, H.: Magnet: a two-pronged defense against adversarial examples. In: ACM Conference on Computer and Communications Security. pp. 135–147 (2017)
11. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: IEEE Symposium on Security and Privacy. pp. 582–597 (2016)
12. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In: MICCAI. pp. 493–501 (2018)
13. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in mri images. IEEE Trans. Med. Imaging **35**(5), 1240–1251 (2016)
14. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014)
15. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. arXiv:1705.07204 (2017)