2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
April 3-7, 2020, Iowa City, Iowa, USA

ROBUST DETECTION OF ADVERSARIAL ATTACKS ON MEDICAL IMAGES

Xin Li

Wayne State University
Department of Computer Science
5057 Woodward Ave., Detroit, MI 48202
xinlee@wayne.edu

Dongxiao Zhu

Wayne State University
Department of Computer Science
5057 Woodward Ave., Detroit, MI 48202
dzhu@wayne.edu

ABSTRACT

Although deep learning systems trained on medical images have shown state-of-the-art performance in many clinical prediction tasks, recent studies demonstrate that these systems can be fooled by carefully crafted adversarial images. It has raised concerns on the practical deployment of deep learning based medical image classification systems. To tackle this problem, we propose an unsupervised learning approach to detect adversarial attacks on medical images. Our approach is capable of detecting a wide range of adversarial attacks without knowing the attackers nor sacrificing the classification performance. More importantly, our approach can be easily embedded into any deep learning-based medical imaging system as a module to improve the system's robustness. Experiments on a public chest X-ray dataset demonstrate the strong performance of our approach in defending adversarial attacks under both white-box and black-box settings.

Index Terms— Adversarial attacks, Medical images, Deep learning, Lung disease classification

1. INTRODUCTION

With the development of deep learning algorithms and the availability of high quality labeled medical imaging datasets, deep learning based medical imaging systems have substantially increased the accuracy and efficiency of the clinical prediction tasks. For example, Daniels et al. [1] extract features from X-rays for lung disease classification, Shaffie et al. [2] detect lung cancer using computed tomography (CT) scans and Reda et al. [3] make an early diagnosis of prostate cancer using magnetic resonance imaging (MRI) scans. Recently, several healthcare start-ups such as Zebra Medical Vision and Aidoc announced U.S. Food & Drug Administration (FDA) clearances for their AI medical imaging systems[1]. These FDA approvals indicate that deep learning based medical imaging systems are potentially applicable for clinical diagnosis in the near future.

[1] https://www.mobihealthnews.com/content/north-america/aidoc-zebra-medical-vision-announce-510k-clearances-ai-image-analysis-software

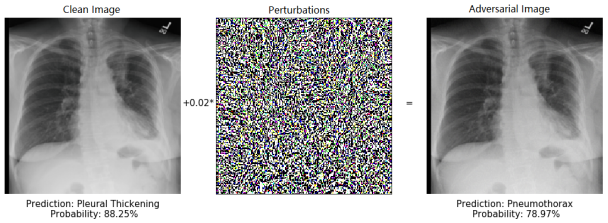

Fig. 1. An adversarial attack against a medical image classifier with perturbations generated using FGSM [4].

In parallel to the progress in deep learning based medical imaging systems, the so-called adversarial images have exposed vulnerabilities of these systems in different clinical domains [5]. Adversarial images are inputs of deep learning models that are intentionally crafted to fool image classification models. Figure 1 shows how a clean image is manipulated to attack a medical image classification system. With only imperceptibly small perturbations added to a clean X-ray image, the system incorrectly classifies "Pleural Thickening" as "Pneumothorax". Consequently, without proper safeguards, users of such systems can be exposed to unforeseen hazardous situations, such as diagnostic errors, medical reimbursement fraud and so on. Therefore, an effective defense strategy needs to be implemented before these systems can be safely deployed.

In response to the threat, several defensive techniques have been proposed. One common strategy in natural imaging domain is adversarial training, which enlarges the training dataset with adversarial images to improve the robustness of the trained Convolutional Neural Network (CNN) model. However, this strategy is not perfect for medical imaging datasets since a large number of diverse adversarial images injected into training dataset can significantly compromise the classification accuracy. To tackle this problem, Ma et al. [6] build a logistic regression classifier based on features extracted from a trained CNN model to discriminate adversarial images from clean images. However, the effectiveness of this approach is restricted to the selected pre-defined attack

methods. To overcome these limitations, Taghanaki et al. [7] equip CNN models with a radial basis mapping kernel, which transforms features onto a linearly well-separated manifold to improve the class separation and reduce the influence of perturbations. He et al. [8] discover that global dependencies and contextual information can be used to improve robustness. Thus, they propose a non-local context encoder in medical image segmentation systems to defend against adversarial attacks. Although both methods increase the robustness by modifying the network architecture, performance of the system may be compromised by the trade-off between accuracy and robustness [9] in practice.

In this paper, we propose a robust detection strategy for adversarial images that can effectively thwart the adversarial attacks against deep learning based medical image classification systems. Inspired by [10], we focus on unsupervised abnormal detection using features extracted from a trained CNN classifier. Our approach does not assume any prior knowledge of attack methods, hence it can robustly defend against diverse unseen attacks, white-box or black-box. Furthermore, our defense strategy can be easily incorporated in any medical imaging system without modifying the architecture nor compromising the performance. Thus it is sufficiently flexible for a wide range of medical imaging problems with various image formats. We use extensive experiments on a public X-ray dataset to demonstrate the effectiveness of our proposed defense approach.

2. METHODS

2.1. Motivation

The adversarial image is crafted by adding subtle perturbations to the original image; as a result, the perturbations at pixel level look like noise which do not impede human recognition. However such noise is obvious at feature levels of CNN models. We demonstrate these characteristics of adversarial medical images by visualizing the feature maps of a CNN model. In Figure 2a, given one clean X-ray image (top left) and its adversarial counterpart (top right), the corresponding feature maps extracted from the first block of a DenseNet-121 [11] are shown in bottom left and bottom right, respectively. It suggests that adversarial perturbations, albeit subtle at pixel level and hard to be detected by human eyes, lead to substantial "noise" at feature levels.

Furthermore, this "noise" can be exacerbated by the convolution-pooling operations implemented in CNN models during forward propagation [12], and finally leads to misclassification. On the other hand, since the magnitude of perturbations increases layer by layer, the clean and adversarial images can be easily distinguished based on the high-level features. This assumption is verified from Figure 2b, which visualizes feature distributions of clean and adversarial X-ray images extracted from the final fully connected layer of the DenseNet-121 using t-SNE method. All X-ray images are randomly selected, which cover different types of pathologies. It is obvious that the clean images can be modeled as a unimodal multivariate density (green) whereas adversarial images (red) can be treated as outliers. Different from natural images that may be affected by changes in lighting and position, medical images are highly standardized since they are generally captured with pre-defined and well-established positioning and exposure. Consequently, the trained deep learning based imaging system is more sensitive to these crafted perturbations.

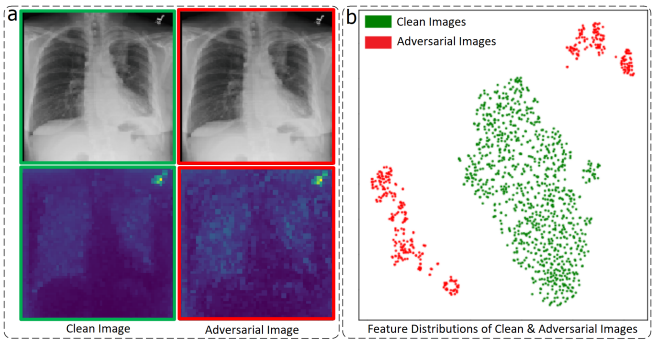

Fig. 2. (a) Visualization of input images and feature maps from the first block of a DenseNet-121 [11]. (b) Visualization of feature distributions from the final fully connected layer of clean X-ray images (green) versus adversarial X-ray images (red).

2.2. Framework

We propose to augment the medical image classification system with an adversarial image detection module. Figure 3 illustrates an example framework of the chest X-ray disease classification system equipped with our detection module. After training the CNN classifier with all clean images to extract the high-level features for learning the detection module, the lower panel illustrates the process of detection and testing. Given a new (clean or adversarial) image, the system extracts features using the trained CNN classifier as the input of detection module. The input image is rejected if detected as an adversarial image, otherwise, it continues to the loss layer to predict classification labels. To accommodate diverse adversarial attacks, we use unsupervised anomaly detection techniques for the detection module. Specifically, we use unimodal multivariate Gaussian model (MGM) as the attacker detection method whereas Isolation Forest (ISO) [13] and One-class SVM (OCSVM) [14] as competing methods.

The high-level feature distribution of clean images can be modeled using MGM: $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{y} = H(\mathbf{x})$ represents the feature extracted using the final fully connected layer given a clean input image $\mathbf{x}$. The $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ are

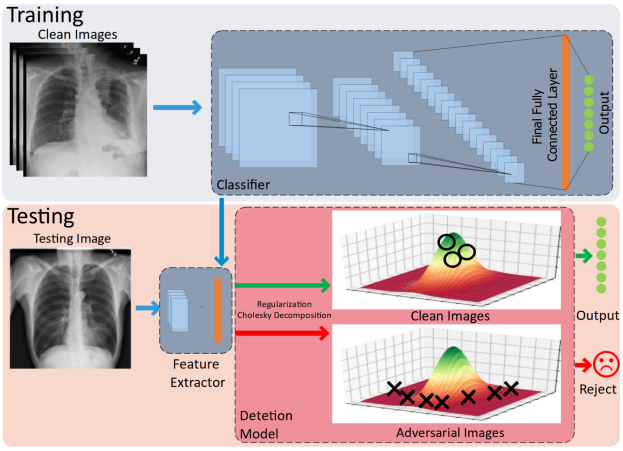

Fig. 3. The proposed defense framework for a chest X-ray disease classification system equipped with our MGM detection module.

mean vector and covariance matrix, where d denotes dimension of MGM. Given features extracted from clean training images $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, we estimate $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$ and $\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T + \lambda\boldsymbol{I}$, where $\lambda\boldsymbol{I}$ is the non-negative regularization added to the diagonal of covariance matrix.

After training MGM, for a new (clean or adversarial) image $\mathbf{x}^*$, we compute the probability of $\mathbf{y}^* = H(\mathbf{x}^*)$ belonging to the clean image distribution by: $p(\mathbf{y}^*) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{y}^* - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}^* - \boldsymbol{\mu})\right)$. However, in practice, the high dimension, i.e., $d = 1024$, makes $p(\mathbf{y}^*)$ computational expensive, and the value of $p(\mathbf{y}^*)$ is so close to zero that cause arithmetic underflow. To overcome these technical difficulties, we use Cholesky decomposition to re-parametrize the covariance matrix: $\boldsymbol{\Sigma} = \boldsymbol{R}\boldsymbol{R}^T$ and rewrite the probability density function into log form: $\log p(\mathbf{y}^*) = -\frac{1}{2}[2\times(\sum_{i=1}^{d}\boldsymbol{R}_{ii}) + \|\boldsymbol{R}^{-1}(\mathbf{y}^* - \boldsymbol{\mu})\|^2 + d\log(2\pi)]$. Finally, as shown in the Figure 3, $\mathbf{x}^*$ will be detected as an adversarial image and rejected if $\log p(\mathbf{y}^*)$ is lower than a threshold. The threshold can be determined by keeping 95% of the training data as clean images.

ISO algorithm builds an isolation tree (itree) by recursively dividing $\mathbf{Y}$ with a random feature and a random cut-off value. By creating many itrees, the average path length of unsuccessful search $c(n)$ is used to assign the anomaly score: $s(\mathbf{y}, n) = 2^{(-E(h(\mathbf{y}))/c(n))}$, where $E(h(\mathbf{y}))$ is the average path length of a single input $\mathbf{y}$. The new (clean or adversarial) image $\mathbf{x}^*$ is rejected if $s(\mathbf{y}^*, n)$ is close to 1. OCSVM is another competitor used in our experiment, it can be summarized as mapping the clean training data $\mathbf{y}$ to a feature space and finding the maximal margin which separates the mapped data from the origin. In our context, let Φ to be the kernel function that transforms $\mathbf{y}$ to another space, and $\mathbf{w}$ and ρ are the parameters to be learned to characterize the maximal margin. After training, given a new (clean or adversarial) image $\mathbf{x}^*$, it will be detected as an adversarial image if the decision function $f(\mathbf{y}^*) = \mathrm{sgn}((\mathbf{w}\cdot\Phi(\mathbf{y}^*)) - \rho) = -1$.

3. EXPERIMENTS

3.1. Settings

Dataset To verify the performance of our proposed defense approach on medical image classification, experiments are conducted on a large public chest X-ray dataset. The NIH ChestX-ray14 [15] contains 112,120 frontal-view chest X-rays taken from 30,805 patients, where around 46% images are labeled with at least one of 14 pathologies. Following the pre-processing of [15], we split the dataset into training, validation and testing datasets by a ratio of 7:1:2 for the image classification system which is DenseNet-121 in our experiment. The features extracted from the entire clean training and validation datasets are used for training and validating the detection module. We then randomly select 1000 clean images from the testing dataset for crafting adversarial images using four adversarial attack methods, i.e., fast gradient sign method (FGSM) [4], projected gradient descent (PGD) [16], basic iterative method (BIM) [17], and momentum iterative method (MIM) [18] (the winner of NIPS 2017 adversarial attacks competition). For each attack method, we craft 1000 adversarial images based on the 1000 clean images.

Attacks We evaluate our defense approaches (MGM, ISO and SVM) against the four attack methods mentioned above. Two attack settings are used in the experiment. 1) White-box Attack: attackers know all details of the true CNN classifier (DenseNet-121), and directly use gradients from the model to craft adversarial images. 2) Black-box Attack: attackers know nothing about the true CNN classifier and use an arbitrary substitute classifier (ResNet-50 [19]) to craft adversarial images. Since the disease classification problem is a multiple binary classification problem and attackers would not know the true label, for each clean image, we use the class with the highest predicted probability to craft the adversarial images. The perturbations are calculated by using the gradient of cross-entropy loss function on the selected class. To ensure the perturbations are subtle enough to remain undetectable from human recognition, the maximum perturbation is limited by 0.05 for black-box setting and 0.02 for white-box setting.

Metrics We evaluate our defense approach against each attack method based on detection performance and follow-up classification performance. The detection performance is evaluated by F1 score, representing the best trade-off between precision and recall. For comparing performance of the follow-up classification, we use AUROC weighted average from 14 different classes because ROC curve has the advantage of determining the optimal cut off values for classification decisions based on the class probabilities. [2]

[2]Code is available at https://github.com/xinli0928/MGM

White-box Attack (F1 / AUROC ± STD)				
Attacks	FGSM	BIM	PGD	MIM
No Defense	0.500 / 0.702 ± 0.063	0.500 / 0.617 ± 0.071	0.500 / 0.616 ± 0.071	0.500 / 0.591 ± 0.063
ISO	0.838 / 0.786 ± 0.077	0.874 / 0.810 ± 0.083	0.874 / 0.810 ± 0.083	0.874 / 0.810 ± 0.083
SVM	0.870 / 0.783 ± 0.077	0.931 / 0.816 ± 0.083	0.931 / 0.816 ± 0.089	0.931 / 0.816 ± 0.094
MGM	**0.936 / 0.801 ± 0.089**	**0.975 / 0.820 ± 0.089**	**0.975 / 0.820 ± 0.089**	**0.975 / 0.820 ± 0.089**
Black-box Attack (F1 / AUROC ± STD)				
Attacks	FGSM	BIM	PGD	MIM
No Defense	0.500 / 0.749 ± 0.077	0.500 / 0.737 ± 0.077	0.500 / 0.741 ± 0.077	0.500 / 0.719 ± 0.077
ISO	0.871 / 0.810 ± 0.083	0.759 / 0.777 ± 0.089	0.735 / 0.776 ± 0.089	0.837 / 0.801 ± 0.083
SVM	0.903 / 0.812 ± 0.077	0.777 / 0.781 ± 0.083	0.754 / 0.776 ± 0.083	0.859 / 0.792 ± 0.089
MGM	**0.958 / 0.819 ± 0.083**	**0.924 / 0.809 ± 0.089**	**0.903 / 0.808 ± 0.083**	**0.957 / 0.818 ± 0.083**

Table 1. F1 scores are shown for comparing detection performance and AUROC values weighted average over 14 different classes with standard deviation are shown for comparing classification performance of each attack-defense combination.

3.2. Results

Table 1 shows the detection performance for each attack-defense combination under both white-box and black-box settings. Since the testing dataset consists of 1000 clean images and 1000 adversarial images, the F1 score of the classification system without a detection module (the weak baseline) is always 0.5. All detection methods demonstrate robust performance against these attacks under the white-box setting with MGM has the best performance. We note that the adversarial images crafted using one-step FGSM [4], an earlier adversarial attack method, are more effective compared to others under the white-box setting evident by a lower F-1 score. Similar to the white-box setting, MGM demonstrates the best performance among all detection methods against all attacks under the black-box setting where the architecture of the true CNN classifier is unknown to the attackers. However, the trend is reversed under the black-box setting that adversarial images crafted using one-step FGSM are easier to be detected compared to others. We explain this phenomenon below.

Since detection is based on the features extracted from the true CNN classifier, an adversarial image is easier to be detected if it is contaminated with more "noise" at feature levels. Under the white-box setting, adversarial images crafted from the iterative methods (e.g., BIM, PGD, MIM) are easier to be detected because they iteratively increase perturbations to maximize the "noise" at feature levels. However, under the black-box setting, adversarial images are crafted using a substitute classifier (ResNet-50), which can be quite different from the true CNN classifier (DenseNet-121). Thus adversarial images crafted by the iterative methods can maximize "noise" for the substitute classifier but not for the true CNN classifier, making it much lower "noise" at feature levels thus harder to be detected.

We also report the follow-up classification performance in Table 1 under both white-box and black-box settings, which is consistent with detection performance. The system equipped with the MGM detection module has the best performance among all detection methods under both settings evident by the highest AUROC values. It is interesting to point out that the proposed framework with a detection module, such as MGM under the white-box setting, can has a better classification performance on mixed clean and adversarial images (0.820) than the true CNN classifier tested only on clean images (0.817), which is possibly due to: (1) the detection module effectively rejects all adversarial images, ensuring the system's non-compromised classification performance as using a clean dataset, and (2) the detection module can also erroneously reject some clean images as adversarial images. These clean images can be problematic for the CNN classifier since they are at tails of the distribution. Therefore, rejecting these clean images can improve classification performance.

4. CONCLUSION

In this paper, we propose an adversarial image detection module for medical imaging classification systems by modeling high-level features learned from clean images using a standard CNN classifier. This strategy does not need any prior knowledge of attack methods nor modification of the CNN architecture. We evaluate the performance of our method under both white-box and black-box settings using a benchmark chest X-ray dataset. This effective strategy can be combined with other defense methods and is sufficiently flexible for many medical imaging applications with diverse image formats. We expect deployment of our approach would enhance the security of deep learning based medical imaging classification systems. For future works, we plan to extend the current method to accommodate more complex datasets that may follow multimodal distributions, and investigate new dimension reduction approaches to reduce the number of training examples required to estimate the distribution.

5. REFERENCES

[1] Zachary A Daniels and Dimitris N Metaxas, "Exploiting visual and report-based information for chest x-ray analysis by jointly learning visual classifiers and topic models," in *ISBI*, 2019.

[2] Ahmed Shaffie et al., "Radiomic-based framework for early diagnosis of lung cancer," in *ISBI*. IEEE, 2019, pp. 1293–1297.

[3] Islam Reda et al., "A new cnn-based system for early diagnosis of prostate cancer," in *ISBI*. IEEE, 2018, pp. 207–210.

[4] Ian J Goodfellow et al., "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[5] Samuel G Finlayson et al., "Adversarial attacks against medical deep learning systems," *arXiv preprint arXiv:1804.05296*, 2018.

[6] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *arXiv preprint arXiv:1907.10456*, 2019.

[7] Saeid Asgari Taghanaki et al., "Vulnerability analysis of chest x-ray image classification against adversarial attacks," in *MICCAI*, pp. 87–94. Springer, 2018.

[8] Xiang He et al., "Non-local context encoder: Robust biomedical image segmentation against adversarial attacks," in *AAAI*, 2019, vol. 33, pp. 8417–8424.

[9] Hongyang Zhang et al., "Theoretically principled trade-off between robustness and accuracy," *arXiv preprint arXiv:1901.08573*, 2019.

[10] Zhihao Zheng and Pengyu Hong, "Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks," in *NIPS*, 2018, pp. 7913–7922.

[11] Gao Huang et al., "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.

[12] Cihang Xie et al., "Feature denoising for improving adversarial robustness," in *CVPR*, 2019, pp. 501–509.

[13] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[14] Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt, "Support vector method for novelty detection," in *Advances in neural information processing systems*, 2000, pp. 582–588.

[15] Xiaosong Wang et al., "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR*, 2017, pp. 2097–2106.

[16] Aleksander Madry et al., "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[17] Alexey Kurakin et al., "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[18] Yinpeng Dong et al., "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193.

[19] Kaiming He et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.