2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Inception-GAN for Semi-supervised Detection of Pneumonia in Chest X-rays

Saman Motamed[1] and Farzad Khalvati[2]

Abstract— **Recent advances in Deep Learning have led to the development of supervised models to detect anomalies in medical images such as pneumonia in chest X-rays. Automatic detection of such anomalies can help clinicians with faster decision making and treatment planning for patients. Nonetheless, supervised models require complete labeled training data with all possible labels (i.e., positive and negative), which are cumbersome and expensive to obtain. We propose an adversarial learning-based semi-supervised algorithm for anomaly detection, which requires training data only with a single class (positive or negative). We applied our proposed Generative Adversarial Network architecture to detect anomalies and score pneumonia in chest X-rays and achieved statistically significant improvements compared to previous state-of-the-art generative network and one-class classifiers for anomaly detection.**

I. Introduction

Pneumonia is a global cause of illness and mortality amongst children under the age of 5 [1]. It has been shown that early diagnosis and detection of pneumonia can minimize the risk factors of the illness [2]. Computer-aided diagnosis using medical imaging has been accelerated over the past decade due to the breakthroughs in the field of Machine Learning and the development of detection and classification architectures that are based on Convolutional Neural Networks (CNNs) [3], [4], [5]. CNNs which are usually used in supervised frameworks, require large amounts of labeled data to automate the task of anomaly detection, such as detecting pneumonia in chest X-rays. Supervised architectures require training data with complete labels for both positive (diseased) and negative (healthy) cases. Nevertheless, this requires accurate labeling of the data for both positive and negative cases. The cumbersome annotation effort and the diagnosis variation amongst expert radiologists limit the performance of supervised models on new (unseen) data. In contrast, solutions based on semi-supervised learning only require partially labeled training data. Semi-supervised learning significantly reduces the cost of creating training data and thus, opens new opportunities for automated disease detection using training data with only single class labels.

Although CNNs have been used extensively in diagnostic medical imaging for disease detection, there is limited work in this regard using semi-supervised architectures such as Generative Adversarial Networks (GANs). Specifically for pneumonia detection using supervised CNNs, Rajpurkar *et al.* [6] used a 121-layer CNN with over 100,000 labeled training data for detection of pneumonia and thirteen other pathologies. They achieved an area under receiver operating characteristic curve (AUC) of 0.76. The only GAN-based model (AnoGAN) was proposed by Schlegl *et al.*, for anomaly detection of retina using OCT images [7].

In this study, we trained the AnoGAN model for detection of pneumonia in X-ray images and proposed modifications to the AnoGAN's architecture to improve the detection of pneumonia from healthy images. The idea behind anomaly detection using GANs comes from the great ability of generative models in learning the image-space manifold where training images lie on, and being able to generate never-before-seen images that lie on the learned image-space [8]. Anomaly detection may be seen as only detecting abnormality in medical images such as a tumour or pneumonia. We extend the definition of anomaly in medical images as the deviation from the image-space manifold of training data. In other words, if the training data only includes healthy (negative) cases, the anomaly detected in the test cases is indeed an abnormality such as tumour. On the other hand, if the training data only includes unhealthy (positive) cases (e.g., images with tumour or pneumonia), the "anomaly" detected in the test cases are the deviation from unhealthy cases meaning that the test case does not contain the disease (i.e., healthy).

The choice for training data between positive or negative cases depends on the availability of a larger amount of data for one class of labels (e.g., positive) compared to the other and the problem of generalizability in training GANs for classification of images. For example, in prostate cancer, radiologists can identify healthy cases with a very high accuracy (over 92%) without a need for biopsy [9], which is both painful and potentially harmful. In this scenario, the proposed GAN-based model can be trained using only negative (healthy) cases. On the other hand, when the radiologists classify prostate cancer patients with high likelihood of cancer, it has been shown that almost all patients indeed have cancer, once biopsied [10]. In this case, the proposed GAN-based model can be trained using only positive (cancerous) cases. In general, highly unbalanced datasets lead to poor performance for supervised models. Thus, it is beneficial to have a semi-supervised model that can be trained using only one class labels with the largest data size. Semi-supervised anomaly detection in medical images has not been explored in domains where anomalies are difficult to detect even for the expert radiologist. This is due to the difficulty of training

[1]S. Motamed is a student at the Institute of Medical Science, University of Toronto, the Hospital for Sick Children, Toronto, Canada sam.motamed@mail.utoronto.ca

[2]F. Khalvati is with the Department of Medical Imaging and Mechanical and Industrial Engineering and Institute of Medical Science, University of Toronto, The Hospital for Sick Children, Toronto, Canada farzad.khalvati@utoronto.ca

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

models like GANs as they get deeper in order to extract more features from images [11]. In this study, we proposed a novel GAN architecture, based on AnoGAN, for the detection of pneumonia in chest X-ray images where we improved the performance of anomaly detection without losing the stability in training the GAN. The advantage of using generative models compared to using CNNs for detection of pneumonia (anomalies) is the dependence of CNNs on labeled data whereas our proposed generative model can be trained only on one class (normal or pneumonia) of chest X-rays and learn to detect anomalies without any label for the anomalous data. We assessed the performance of the proposedm odel using an anomaly score measure. We evaluated the current state-of-the-art GAN based AnoGAN [7] and the state-of-the-art one class classifier Deep SVDD [12] against our proposed Inception-GAN and showed statistically significant improved results in detection of pneumonia in chest X-rays.

II. Generative Adversarial Learning for Pneumonia Detection

To identify positive from negative cases, we learn a model representation only on positive (or negative) cases using a GAN. This method trains a generative model (**G**) to learn the representation of positive (or negative) cases and a discriminator model (**D**) to identify real from generated images simultaneously. In our experiments, we use X-ray images to identify patients with pneumonia from healthy ones.

A. Inception-GAN Model

The Generator (G) (Fig. 1) learns a distribution P_g over the input data x via mapping of input noise z, to $2D$ images by function $G(z)$. The trained Generator learns the mapping $G(z) : z \longmapsto x$ from latent space representations z to realistic, $2D$, X-ray images. Our Generator model follows DCGAN's (AnoGAN and DCGAN follow the same architecture) architecture [13] with two main modifications; the use of Inception and Residual Blocks, as shown in Fig. 2.

The idea behind the Inception and residual architecture [14] is being able to increase GAN's ability to capture more details from training image-space without losing spatial information after each convolution and pooling layer. Although making the Generator deeper is theoretically a valid way to capture more long-range details in the image, deep GANs are unstable and hard to train [13], [11].

The Discriminator (**D**) is a 4-layer CNN that maps a 2D image to a scalar output that can be interpreted as the probability of the given input being a real chest x-ray sampled from training data or generated G(z) by the Generator G.

During training, Generator G is trained to minimize the accuracy of Discriminator D's ability in distinguishing between real and generated images while the Discriminator is trying to maximize the probability of assigning real training images the "real" and generated images from G, "fake" labels. The Generator improves at generating more realistic images while Discriminator gets better at correctly identifying between real and generated images.

III. Experiments

A. Data and Pre-processing

We used the publicly available chest X-ray images for children [15] with two categories of Normal, $1,575$ images, and Pneumonia, $4,265$ images. The images were in jpeg format and varied in size, with pixel values in the [0, 1] range. We resized all images to 128×128 pixels. Images were normalized to have [-1, 1] range for the purpose of *tanh* non-linearity activation in our GAN architecture. Given that the positive (pneumonia) cases are much larger than the negative cases (4,265 vs. 1,575), we chose to train the proposed architecture with positive cases and then tested on both positive and negative cases. We split pneumonia images into 3,765 training and 500 test images. To keep the test set balanced, we randomly chose 500 images of normal cases and added them to our test data yielding a total of 3,875 training and 1,000 test images.

B. Competing Methods

Ruff *et. al* proposed a Deep One-class classification model (Deep SVDD) [12] that outperformed shallow and deep semi-supervised anomaly detection models at the time, including AnoGAN. We compare our Inception-GAN against these models as baselines.

C. Shallow Baselines

We followed the same implementation details of the shallow models as used in Ruff *et. al*'s Deep SVDD study. (i) **One-class SVM** (OC-SVM) [16] finds a maximum margin hyper-plane that best separates the mapped data from the origin. (ii) **Isolation Forest** [17] (IF) isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. We set the number of trees to t = 100 and the sub-sampling size to 256, as recommended in the original work

D. Deep Baselines

Our Inception-GAN is compared with two deep approaches. (i) Ruff *et. al*'s **Deep SVDD** showed improved accuracy of one class classification in a framework where one class from MNIST [18] and CIFAR-10 [19] was kept as the known image, and the rest of the classes were treated as the anomaly. Deep SVDD learns a neural network transformation from inputs into a hypersphere characterized by center *c* and radius *R* of minimum volume. The idea is that this allows for the known (pneumonia) class of images to fall into the hypersphere and the unknown (healthy) class to fall outside of the hypersphere. (ii) **AnoGAN** is trained as the base GAN benchmark for the task of pneumonia detection [13].

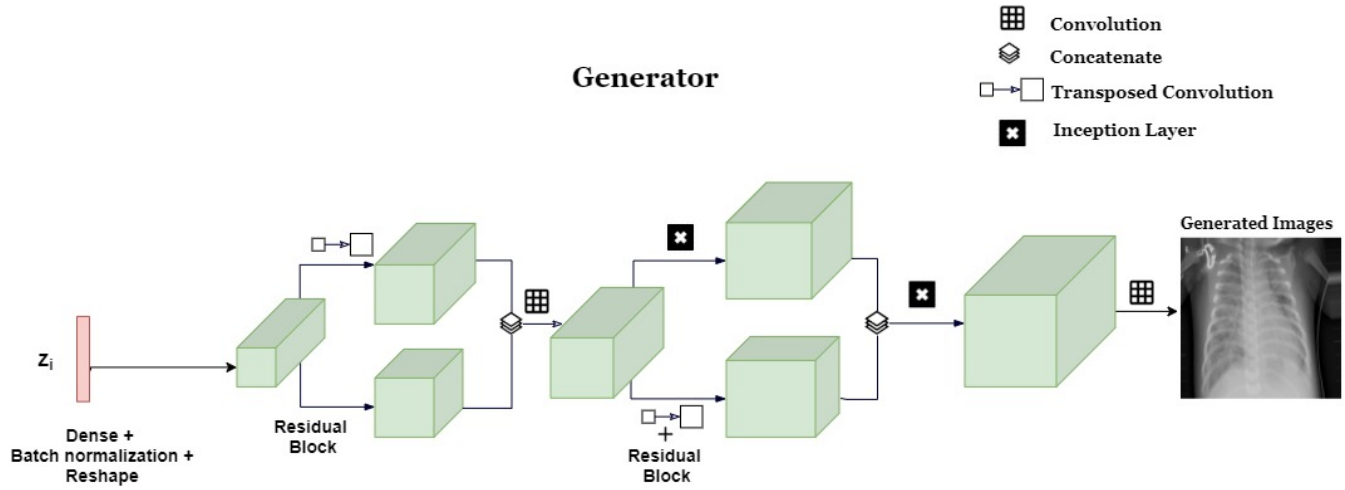

Fig. 1. Generator Architecture

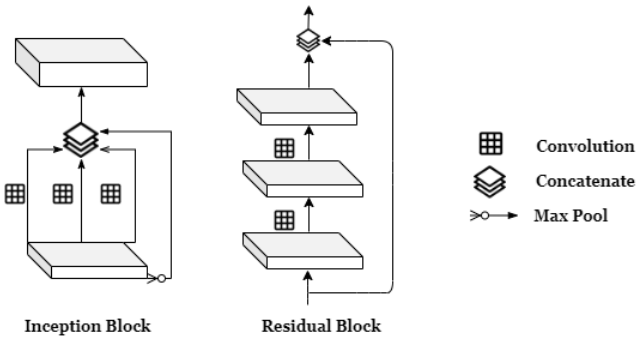

Fig. 2. Inception and Residual Architecture

E. Evaluation

When Inception-GAN is trained, the Generator has learned the mapping $G(z) : z \longmapsto x$ from latent space representation z to realistic images (Chest x-ray with pneumonia). Given a query image x in the test, we want to find a point z from the latent space that, given the Generator's output on that point, $G(z)$, is most similar to the query image x. The expected behaviour after successful training is that the query image x, if affected by pneumonia, will result in finding an image $G(z)$, which is visually closer to image x than if the query image was a normal case.

To find latent variable z that generates the most similar image $G(z)$ to the query image x, we use backpropagation with a predefined number of steps. The loss function defined to find such z through backpropagation is comprised of two components; *residual loss* and *discrimination loss*. Residual loss ($\mathcal{L}_\mathcal{R}$) calculates the L1 distance between $G(z)$ and the query image x and enforces visual similarity between the query image and generated image.

$$\mathcal{L}_R(z_i) = \sum |x - G(z_i)| \tag{1}$$

Schlegl *et al.* [7] proposed a discrimination loss ($\mathcal{L}_\mathcal{D}$) inspired by the concept of feature matching [20] that enforces generated images $G(z_i)$ to follow the statistical characteristics of the training images. $\mathcal{L}_\mathcal{D}$ is defined below where the output of an intermediate layer of the discriminator, $f(.)$, is used to represent the statistical characteristics of the input image.

$$\mathcal{L}_D(z_i) = \sum |f(x) - f(G(z_i))| \tag{2}$$

The overall loss used to backpropagate and find the best z is a weighted sum of residual and discrimination loss;

$$\mathcal{L}(z_i) = (1 - \lambda) \times \mathcal{L}_R(z_i) + \lambda \times \mathcal{L}_D(z_i) \tag{3}$$

The Anomaly score $A(x)$ for the query image x is defined as;

$$A(x) = (1 - \lambda) \times R(x) + \lambda \times D(x) \tag{4}$$

where R(x) and D(x) are respectively the residual and discrimination loss of the best z_i found through backpropagation. λ adjusts the weighted sum of the overall loss and anomaly score. We used $\lambda = 0.2$ to train our proposed Inception-GAN and DCGAN [7]. Both architectures were trained with the same initial conditions for performance comparison.

IV. Results

We computed anomaly score $A(x)$ (eq. **IF**, **OC-SVM** and **Deep SVDD**, trained on the pneumonia cohort, generated an anomaly score based on their objectives (e.g, Deep SVDD calculates the distance between data to center of hypersphere as anomaly score). We compared the performance of Inception-GAN to shallow and deep competing models by calculating the area under the ROC curve (AUC). Table I shows the AUC of each model, achieved on 1,000 test cases where Inception-GAN (AUC of 89%) outperformed all competing models with the previous state-of-the-art GAN-based model, AnoGAN, achieving the second best performance for detecting pneumonia from healthy X-rays. DeLong test [21] was used to compare the AUC of the models by calculating the *p-value* for significance difference. DeLong test showed significant improvement of Inception-GAN over AnoGAN

and Deep SVDD with a $p = 0.01$ and $p = 2.4 \times 10^{-3}$ respectively.

OC-SVM	IF	AnoGAN	Deep SVDD	Inception-GAN
0.66	0.64	0.87	0.86	**0.89**

TABLE I
ONE-CLASS CLASSIFICATION AUCS FOR CLASSIFYING PNEUMONIA AND NORMAL X-RAYS

V. CONCLUSION

In this paper, we proposed an adversarial learning model, Inception-GAN, that improves on the state-of-the-art for the task of pneumonia classification in chest X-ray images. GANs have the advantage over supervised methods by eliminating the need for labeled data in order to locate and detect anomalies as long as there are enough cases of one class (healthy or diseased) for the model to learn from. Our study had the limitation of having access to more abnormal data than normal images. In order to achieve reliable results, we reversed the learning strategy to learn "abnormal" cases in training instead. The underlying concept, however, remains the same. As future work, we will apply the proposed architecture to cancer detection and localization with no need for tumour location in training data.

VI. ACKNOWLEDGEMENTS

This research was funded by Chair in Medical Imaging and Artificial Intelligence funding, a joint Hospital-University Chair between the University of Toronto, The Hospital for Sick Children, and the SickKids Foundation.

REFERENCES

[1] TKP Nguyen, TH Tran, CL Roberts, GJ Fox, SM Graham, and BJ Marais. Risk factors for child pneumonia-focus on the western pacific region. *Paediatric respiratory reviews*, 21:95–101, 2017.

[2] Jose Manuel Pereira, Jose Artur Paiva, and Jordi Rello. Severe sepsis in community-acquired pneumonia—early recognition and treatment. *European journal of internal medicine*, 23(5):412–419, 2012.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[6] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[7] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Niranjan J Sathianathen, Badrinath R Konety, Ayman Soubra, Gregory J Metzger, Benjamin Spilseth, Paari Murugan, Christopher J Weight, Maria A Ordonez, and Christopher A Warlick. Which scores need a core? an evaluation of mr-targeted biopsy yield by pirads score across different biopsy indications. *Prostate cancer and prostatic diseases*, 21(4):573–578, 2018.

[10] Darren Foreman, jonathon cho, Michael O'Callaghan, and andrew fuller. How predictive of prostate cancer are pirads 4 or 5 lesions on mri? *World Journal of Urology*, 33:13, 10 2015.

[11] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.

[12] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[13] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[15] Daniel Kermany, Kang Zhang, and Michael Goldbaum. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2, 2018.

[16] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[17] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[18] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[21] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.