Assessment of Data Augmentation Strategies Toward Performance Improvement of Abnormality Classification in Chest Radiographs

Prasanth Ganesan, M.S.*†, Sivaramakrishnan Rajaraman, Ph.D.*, Rodney Long, M.A.*,
Behnaz Ghoraani, Ph.D.†, Sameer Antani, Ph.D.,*‡ *Senior Member, IEEE*

***Abstract*— Image augmentation is a commonly performed technique to prevent class imbalance in datasets to compensate for insufficient training samples, or to prevent model overfitting. Traditional augmentation (TA) techniques include various image transformations, such as rotation, translation, channel splitting, etc. Alternatively, Generative Adversarial Network (GAN), due to its proven ability to synthesize convincingly-realistic images, has been used to perform image augmentation as well. However, it is unclear whether GAN augmentation (GA) strategy provides an advantage over TA for medical image classification tasks. In this paper, we study the usefulness of TA and GA for classifying abnormal chest X-ray (CXR) images. We first trained a progressive-growing GAN (PG-GAN) to synthesize high-resolution CXRs for performing GA. Then, we trained an abnormality classifier using three training sets individually – training set with TA, with GA and with no augmentation (NA). Finally, we analyzed the abnormality classifier's performance for the three training cases, which led to the following conclusions: (1) GAN strategy is not always superior to TA for improving the classifier's performance; (2) in comparison to NA, however, both TA and GA leads to a significant performance improvement; and, (3) increasing the quantity of images in TA and GA strategies also improves the classifier's performance.**

Keywords: Deep learning, Generative adversarial network, Medical image synthesis, Chest X-ray, Abnormality classification, Progressive-growing GAN

I. Introduction

Training datasets play an important role in regulating supervised classifier performance including deep learning networks [1]. Particularly, for medical image training sets there are two major challenges: obtaining sufficient labeled data and obtaining class-balanced data in case of multi-class training. Traditionally, both these challenges have been tackled using image augmentation. Apart from this, image augmentations have also been used for preventing model overfitting by introducing data diversity and regularization. The traditional augmentation (TA) techniques include transformations, such as rotation, translation, channel splitting, Gaussian smoothing, unsharp masking, and etc. These image transformations improve model robustness, generalization, and learn better characteristics for making image distinctions. Generative Adversarial Networks (GANs) are well-known for generating synthetic data close to the training set distribution [2]. For this reason, they have been largely used for medical image synthesis – as a matter of fact, 36% of the GAN-based papers in medical imaging domain pertains to synthesizing images, followed by a segmentation and image reconstruction [3]. Hence, GAN is a powerful tool for image augmentation in medical image datasets.

Although GANs have been used to synthesize MRI, CT [4], and natural image modalities [5], their potential in synthesizing chest X-ray (CXR) images have not been fully explored. Previous works published on CXR GAN augmentations (GA) include assessment of GA for cardiomegaly and multiple chest pathology classification using Deep Convolution GAN (DC-GAN) [6], [7]. The goal of this paper is to study the effectiveness of GA and TA on a CXR abnormality classifier's performance, with the baseline as a training set with no augmentation (NA). We first implement the progressively-growing GAN (PG-GAN) model [8] and train it on a large dataset of normal and abnormal CXR images. Then, we train a CXR abnormality classifier with NA, TA, and GA training sets individually, and study the performance variation of the classifier for each type of augmentation strategy. In addition, we vary the number of augmented images that was included in each training set, and determine if the classifier performance is affected by the quantity of image augmentations. Our analysis show that TA improves the classifier's performance in most cases than GA, however, both TA and GA contributes largely to performance improvement when compared to NA. Hence, this study concludes that image augmentations offer improvement in model performance, but GA strategy needs further investigation on its relative performance degradation, contrary to some works reporting better performance for GA strategy than TA [6], [7]. Further, it can be concluded that the number of augmented images in TA strategy also plays a role in the classifier's performance.

II. Methods

We performed a comparative study between NA, TA and GA strategies for improving the performance of an abnormality classifier for CXR images. The following sections describe how the performance analysis was done and the deep network models used in various parts of the study.

A. PG-GAN Model Overview

PG-GAN was introduced by Karras *et al.,* [8] as an attempt to synthesize high-resolution images (upto 1024x1024 pixels), which were not realizable by DC-GANs. The primary reason for the difficulty in achieving high-resolution images before the introduction of PG-GANs was that the weights

*National Library of Medicine, National Institutes of Health, Bethesda, MD
†Department of Electrical Engineering, Florida Atlantic University, Boca Raton, FL
‡Corresponding author, Email: sameer.antani@nih.gov

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

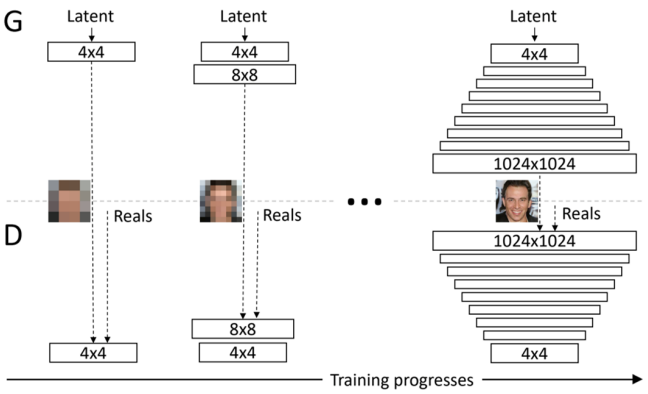

Fig. 1. Generator and Discriminator of PG-GAN – The network is progressively grown starting from 4x4 and trained until it reaches the required resolution. Image source [8]

could not be learned consistently while searching for the global minimum using a network of that size. To circumvent this limitation, the weights in PG-GAN are learned by progressively growing the network starting from a 4x4 up to the final resolution. For example, in case of an image with a resolution of 512x512 pixels, the image is first downsampled into resolutions of 4x4, 8x8, 16x16, and all the way to 512x512. Then, the network is trained starting from the smallest resolution. After training the smallest resolution, the network is grown to the next resolution and the training is continued using the previous weights. In other words, the weights of the smaller resolutions are fully learned before switching the network to the higher resolution, thus making it possible for the network to reach the global minimum for the higher resolution weights. The network model is shown in Figure 1. More details on the PG-GAN model can be found in Ref. [8]. We utilize this model to synthesize high-resolution CXR images as explained in the following section.

B. Dataset and Training Details of PG-GAN

The goal of training a PG-GAN is to perform GA for the abnormality classifier, using realistic CXR images of normal and abnormal classes. The dataset used in this study is made available for the Radiological Society of North America (RSNA) machine learning challenge (https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data) by the joint effort of radiologists from the RSNA and Society of Thoracic Radiology (STR). The dataset includes images with pulmonary opacity that may represent pneumonia or other disorders and other images with no abnormal findings. All the images were of 1024x1024 pixel dimensions with an 8-bit depth. The images were first pre-processed by segmenting the lung region of interest (ROI) using a dropout UNET [9]. This segmentation helps to remove irrelevant regions that carry structures that do not contribute to the abnormality, so that the PG-GAN can focus on learning the ROI. The UNET model consists of dropout layers following a Gaussian distribution, after every pair of convolution and ReLU layers. The addition of Gaussian noise is expected to mimic the noise present during CXR image acquisition. The resulting images were cropped to a bounding box containing the lungs and were then resized to 512x512 pixels. These 512x512 images (8,954 normals and 11,653 abnormals) were used for training the PG-GAN.

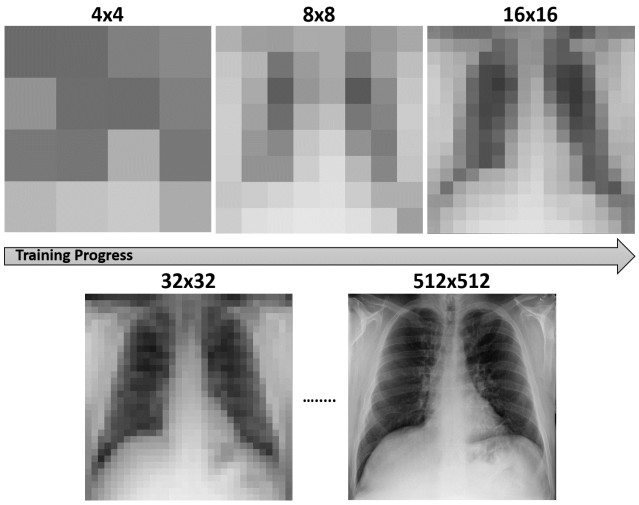

Fig. 2. Random Samples During Training – For each network size, the training progression becomes visually better, and at the final 512 resolution, the sample is at its best quality.

The normal and abnormal images were randomly selected from the RSNA collection, and split into equal number ($N = 6268$) for training set while the rest ($N = 2686$ for normals and $N = 5385$ for abnormals) were used for test set. The PG-GAN was individually trained on the normal and abnormal images in the training set. Each training phase took about six days on a high-performance machine with an NVIDIA GTX 1080Ti GPU and 48GB RAM. An example of progressive resolutions at different training instances is shown in Figure 2. After the network was fully trained, which took about 150 epochs, images of each class were synthesized and used in the GA-training set of the abnormality classifier as described below.

C. Abnormality Classifier Model

Transfer learning assists in faster loss convergence by initializing the current model with the learned weights of a pre-trained model [10]. In this study, a pre-trained VGG16 model [11] was used and customized for CXR abnormality classification. The model was truncated at the last convolution layer and a global average pooling (GAP) followed by a final dense layer was added to output the binary labels. The model architecture is shown in Figure 3.

The model was initialized with ImageNet weights and then fine-tuned end-to-end to learn the hierarchical feature representation from the CXRs. A randomized grid search [12] was performed to obtain the best values for the hyperparameters that include momentum, learning rate, and L2-weight decay. The search range for these parameters were set to [0.8, 0.9],

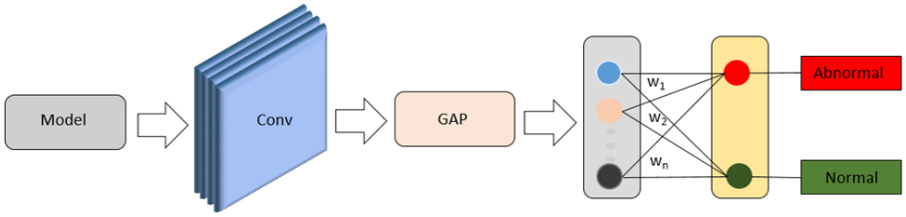

Fig. 3. Abnormality Classifier Architecture – Custom VGG16 model initialized with ImageNet weights for classifying normal and abnormal CXRs. Conv is convolutional layers and GAP is global average pooling.

[1e-6, 1e-3], and [1e-10, 1e-2], respectively. This classifier was then trained using various augmentation strategies as explained in the following section.

D. Training Using NA, TA and GA training sets

1) No Augmentation (NA): The same training set ($N = 6268$ normal and abnormal images each) that was described before was used to train the abnormality classifier without any augmentation. The model was trained for 100 epochs the learning rate was decreased whenever the validation accuracy ceased to improve. Model training took an hour on a Linux machine with 1080Ti GPU and 64GB RAM. All the training described henceforth used the same training set as the base for performing augmentation.

2) Traditional Augmentation (TA): TA was performed on the training set images during the run-time of each mini-batch of the classifier training. As mentioned earlier, we aimed to also study the effect of varying number of augmentations, in addition to the types of augmentation strategies. For this purpose, we split the augmentations into 25%, 50%, 75% and 100% of the number of training samples, and performed the training on each of the augmentation proportion set.

The transformations applied as part of the TA were Gaussian smoothing, unsharp masking, and minimum filtering. Gaussian smoothing helps to reduce the electronic noise due to random variations in the brightness and color and improve the detection of edges present in the image. Unsharp masking helps in amplifying the high-frequency image components, which again pertains to the edges. Minimum filtering finds the minimum of the pixels within a localized region and in turn aids to remove positive outlier noise in images. The model was trained for 100 epochs, and the training took between one to three hours for varying proportion of augmentations to complete on the machine with the same configuration mentioned in NA.

3) GAN Augmentation (GA): Using the trained PG-GAN, CXRs were synthesized and added to the training set by following the same proportions as described above (25% - 100% of images present in the training set). Hence, a total of 6268 (100% of the training set) images of 512x512 pixel dimensions were synthesized for each class. Example of some synthesized images of normal and abnormal classes are shown in Figure 4.

The optimal values of momentum, learning rate, and L2 weight decay were found to be 0.9, 1e-4, and 1e-6,

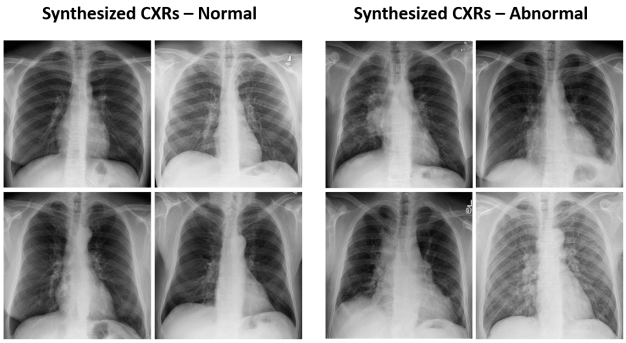

Fig. 4. Synthesized CXRs Examples – Some examples of normal and abnormal CXRs generated by the PG-GAN. The images can be visually deemed as having considerable quality and overall structure.

respectively. The trained models were evaluated with the test set to make predictions. The performance metrics were computed as discussed in the following section.

E. Classifier Performance Analysis

The performance of the abnormality classifier was analyzed based on the following metrics: accuracy (ACC), area under the curve (AUC), F-Score, and Matthew's Correlation Coefficient (MCC). These metrics were computed for NA, TA, and GA cases along with cases of different proportion of augmentations. These results are shown in Table I.
The following observations can be made from Table I:
1. The performance of the classifier with augmentation (both TA and GA) is always higher than the performance of the classifier without augmentation (NA). This observation is coherent with various other studies [13], [5] which also emphasizes that augmentation intends to reduce model overfitting and improve generalization by introducing diversity in the training set distribution, hence improving the classifier performance.

2. The performance of the classifier trained on TA dataset is higher than the one trained on GA dataset. This observation is counter-intuitive due to the fact that the purpose of GA is to generate realistic images that resembles the training set. However, it could be that the lower performance may be attributed to the limited variance exhibited by GA images. In general, data augmentation is performed to introduce controlled variance during training. So, the PG-GAN-generated high-resolution images provide little variance for

TABLE I

SUMMARY OF PERFORMANCE METRICS – PERFORMANCE OF THE ABNORMALITY CLASSIFIER USING NA, TA AND GA TRAINING SETS. BOLD NUMBERS INDICATE BETTER PERFORMANCE.

Method	% Augmentation	ACC	AUC	F-SCORE	MCC
NA (Baseline)	0	79.32	86.03	0.84	0.55
TA	25	**81.20**	**88.01**	**0.86**	**0.58**
	50	**81.38**	**88.34**	**0.86**	**0.59**
	75	**82.23**	**88.57**	**0.86**	**0.59**
	100	**84.25**	**90.70**	**0.88**	**0.64**
GA	25	80.82	87.84	0.85	**0.58**
	50	81.34	88.05	**0.86**	0.58
	75	80.41	87.42	0.85	0.58
	100	83.94	90.29	**0.88**	**0.64**

the abnormality classifier to learn, as opposed to TA, where the augmented data mimics the biological variance present in the unseen test set.

Another perspective could be that with TA strategies, the abnormality classifier may find it relatively easier to approximate the noise functions in TA images as compared to approximating (especially since it has fewer parameters) the images generated by GANs which use large and complex functions for synthesis. However, the improvements in performance between GA and TA are only modest, and further analysis is required to interpret the performance degradation due to GA.

3. The performance of the classifier increases with an increase in the proportion of augmented images. From the performance metrics in Table I, it can be observed that the performance increases gradually as the percentage of the augmentations increase irrespective of the augmentation strategy. Maximum performance is for 100% augmentation, which basically means the number of images augmented is twice the original training set ($N = 6,268 \times 2 = 12,536$). The performance of the classifier for augmentations higher than 100% would be investigated in our future work.

III. CONCLUSION

In this paper, we analyzed the effect of various data augmentation strategies on the performance of abnormality detection of CXR images. The PG-GAN was able to generate visually realistic CXRs, however, the augmentation using GAN-generated images slightly lowered the performance of the abnormality classifier compared to the augmentations using traditional techniques. This might be due to the inability of GANs to expand the information space while adding more samples for training. Further, it is more difficult to train a GAN rather than applying TA to the data. The performance with augmentation compared to that without any augmentations showed improvement regardless of the type of number of augmentations. Also, we found that the performance of the classifier improves with increasing the number of augmentations for both TA and GA strategies. Further investigation is needed to learn why GA degrades the classification performance and also to study the effect of increasing the number of augmentation to $>100\%$ on the performance of the CXR abnormality classifier.

IV. ACKNOWLEDGEMENT

This work was supported by the Intramural Research Program of the Lister Hill National Center for Biomedical Communications (LHNCBC), the National Library of Medicine (NLM), and the U.S. National Institutes of Health (NIH).

REFERENCES

[1] Sivaramakrishnan Rajaraman, Sameer K Antani, Mahdieh Poostchi, Kamolrat Silamut, Md A Hossain, Richard J Maude, Stefan Jaeger, and George R Thoma. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568, 2018.

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

[3] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *arXiv preprint arXiv:1809.06222*, 2018.

[4] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer, 2017.

[5] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[6] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741M. International Society for Optics and Photonics, 2018.

[7] Hojjat Salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 990–994. IEEE, 2018.

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[9] Alexey A Novikov, Dimitrios Lenis, David Major, Jiří Hladvka, Maria Wimmer, and Katja Bühler. Fully convolutional architectures for multi-class segmentation in chest radiographs. *IEEE Transactions on Medical Imaging*, 2018.

[10] Sivaramakrishnan Rajaraman, Sema Candemir, Incheol Kim, George Thoma, and Sameer Antani. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Applied Sciences*, 8(10):1715, 2018.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[13] Jared A Dunnmon, Darvin Yi, Curtis P Langlotz, Christopher Ré, Daniel L Rubin, and Matthew P Lungren. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, page 181422, 2018.