Chest X-ray Disease Diagnosis Using Generative Adversarial Networks

Santhanu Venugopal Sunitha
Georgia Institute of Technology, Atlanta, Georgia, U.S.A

Abstract

Imbalanced datasets are often a major problem in Machine Learning, causing classifiers to be more biased towards majority classes in the dataset. A medical imaging dataset containing less examples of rare diseases can make Machine Learning models to inaccurately predict the presence of the same in patients. This study proposes to simualate and augment CheXpert Chest x-ray imaging dataset[3] using Generative Adversarial Networks techniques for accurate classification of rare diseases, which otherwise could not be effectively diagnosed in an imbalanced dataset and then compare the accuracy of the classifier against the CheXpert baseline model.

The code is available on Github here

Introduction

The chest x-ray is the most commonly performed diagnostic x-ray examination and helps physicians to diagnose various medical conditions like Pneumonia, emphysema and cancer. With the advent of Deep Learning techniques for image classification, Medical imaging datasets of chest x-rays can be effectively used for disease diagnosis without the intervention of a practicing radiologist. However, when it comes to imbalanced datasets, Machine Learning tends to produce unsatisfactory classifiers. In the field of Medical imaging, this problem relates to inadequacy of data relating to rare diseases, hence hampering it's automated diagnosis.

Unsupervised Learning techniques are usually useful in exploratory data analysis as it helps to identify the hidden patterns in the data without using labeled examples. However, in 2014, Generative Adversarial Networks[1] (GAN) was proposed by Ian Goodfellow and other researchers , which being an unsupervised learning technique, could learn to mimic any data distribution. Hence, GAN can be used as a tool for simulation and augmentation of imbalanced data and hence improve the accuracy of the medical image classifier.

The most commonly used benchmark for comparing chest radiograph classifiers has been the ChestXray14 dataset provided by Wang et al[5], which helped in significant development of chest radiology Deep CNN classifers. However, as noted in Rajpurkar et al[4], the labels for this dataset was derived using an automatic labeler, without any radiologist validation. CheXpert dataset[3] on the other hand, comes with labels validated by radiologists and hence will guarantee strong reference standards for building better and accurate image classifiers.

A recent study[2] applied a Deep Convolutional GAN (DCGAN) technique to augment the chest x-ray images of five rare pathological diseases, thereby balancing the dataset and improving the classification performance.This study along with DCGAN, will also focus on other GAN techniques to simulate images of imbalanced classes in CheXpert Dataset[3] and compare performance with classifiers trained on real images and also compare with previous set benchmarks. Another study[8] shows how small imaging datasets can be synthetically augmented using GAN techniques to achieve better accuracy for detecting Liver Lesions.

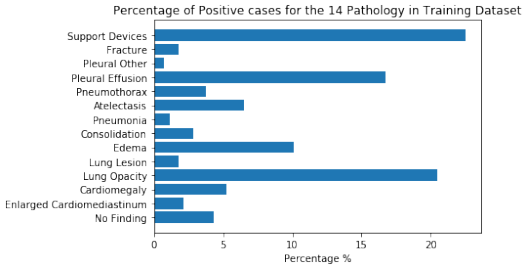
Data and Descriptive Statistics

The CheXpert training dataset consists of 223,415 chest radiographs of 64,540 patients. It consisted of Xray images of 28,728 female patients and 35,811 Male patients with one patient's gender labeled as unknown. The average age for female and male patients were around 61 years and 59 years respectively. The Xray images were labeled for the presence of 14 diseases as either positive (value 1), negative (value 0), uncertain (value -1) or blank (unmentioned).

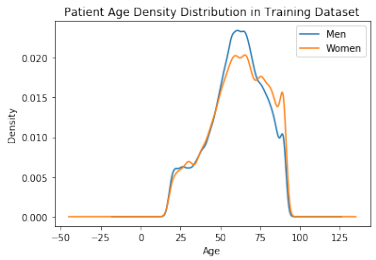
Table1 depicts the count and percentage of the 14 pathological cases in the training dataset.As noted from Figure 1, the CheXpert training dataset doesnot have balanced pathological classes. Figure 1 shows that the majority of male and female patients are between 50 and 75 years of age.

Pathology	Positive	Uncertain	Negative
No Finding	22381 (10%)	0	201033 (90%)
Enlarged Cardiomediastinum	10798 (4.8%)	200213 (89%)	12403 (5.5%)
Cardiomegaly	27000 (12.08%)	188327 (84.3%)	8087 (3.6%)
Lung Opacity	105581 (47.25%)	112235 (50.2%)	5598 (2.5%)
Lung Lesion	9186 (4.1%)	212740 (95.2%)	1488 (0.6%)
Edema	52246 (4.1%)	158184 (70.8%))	12984 (5.8%)
Consolidation	14783 (6.6%)	180889 (80.9%)	27742 (12.4%)
Pneumonia	6039 (2.7%)	198605 (88.9%)	18770 (8.4%)
Atelectasis	33376 (14.9%)	156299 (69.9%)	33739 (16.5%)
Pneumothorax	19448 (8.7%)	200821 (89.8%)	3145 (1.4%)
Pleural Effusion	86187 (38.5%)	125599 (56.2%)	11628 (5.2%)
Pleural Other	3523 (1.5%)	217238 (97.2%)	2653 (1.1%)
Fracture	9040 (4%)	213732 (95.6%)	642 (0.2%)
Support Devices	1079 (0.4%)	106334 (47.5%)	116001 (51.9%)

Table 1: Statistics on Studies Containing the 14 Pathology in the Training Dataset.



(a) Percentage of Positive Pathology in the Dataset



(b) Density Distribution of Male and Female Age

Figure 1: Descriptive Statistics

Approach

The inital approach will use the DenseNet121[3,6] pre-trained models implemented in Pytorch, perform chest x-ray image classification with the CheXpert training dataset using Tranfer Learning and set the benchmarks to improve upon. Then, GAN techniques for data augmentation and simulation will be applied to the trained models using Pytorch-GAN and train the aforementioned ConvNets against the newly generated images.Finally, the final trained models will be evaluated using AUC ROC curve on the validation set and the performances will be compared.

In order to recreate the baseline DenseNet121[3,6] model as implemented by the ChexPert team, which is detailed in the article Irvin, J. et al[3], a Pytorch implementation of the DenseNet121[3,6] architecture with pre-trained weights was applied for the baseline training. A study[7] on chest xray images of lung cancer shows the effectiveness of applying transfer learning for prediction of lung cancer by using a DenseNet121[3,6] model multiple times with the last layer being modified in accordance with their requirement.

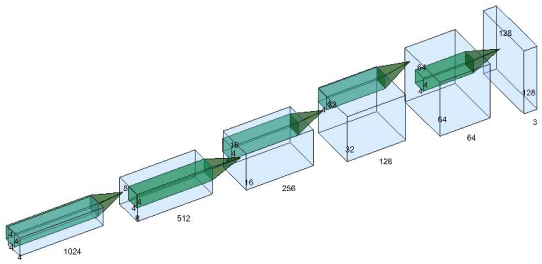
Based on the ChexPert article[3], they have implemented various techniques to deal with the uncertainity labels like U-Zeroes , U-Ones , U-Ignore , U-SelfTrained and U-MultiClass, but the best AUC score was achieved mainly for U-Zeroes and U-Ones approaches. Hence, only these two approaches are chosen to implement the baseline model to impove upon.

In order to generate artificial images through Generative Adversarial networks, a Deep convolutional generative adversarial networks (DCGAN)[9] was trained. The DCGAN was implemented using a pytorch implementation (Source) which was custom tailored for Chexpert Dataset. The generated artifical images were then merged with real images to train the final classifier, followed by drawing comparison between the baseline and my approach.

DCGAN

A basic GAN implemenation consist of a Generator and Discriminator , which in case of DCGAN are CNNs with convolutional-transpose layers and convolutional layers respectively[9] . The generator tries to generate 'fake' images of the Chest Xray by learning the data distribution - p_{data} of the xray images while the discrimator tries to discriminate between real xray images from training set and fake generated images by the Generator. A Nash equilibrium is reached when the discrimnator classifies images as Fake or Real with 50% probability.

The Generator takes a latent space vector - z as input which was sampled from a normal distribution and is fed to strided convolutional-transpose layers and tranformed into 3 x 128 x 128 images, represented by G(z).The discriminator is a binary classifier , which outputs a value D(G(z)) to predict the chance that the input from Generator is real or not.

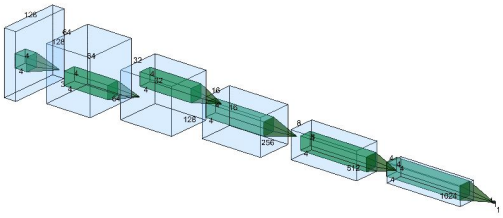


(a) Generator Architecture used to create 128 x 128 images from the latent space vector z

Figure 2: DCGAN Generator Architecture (Created with Source)

The GAN loss function is defined as below in equation , which basically is a cross entropy loss function. During training, the discriminator tries to maximize the below mimimax equation as it wants to classify real samples from fake ones accurately. However, the generator tries to maximize it's output G(z) and tires to trick the discriminator to minimize 1-D(G(z)). Two gradient updates are made to update the parmeters of discriminator and generator with respect to the loss functions. The convergence is defined at $p_{data} = p_z$, though that is usually hard to achieve.

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}\big[logD(x)\big] + E_{z\sim p_z(z)}\big[log(1-D(G(z)))\big]$$



(a) Discriminator Architecture used to classify the 128 x 128 generated image

Figure 3: DCGAN Discriminator Architecture (Created with Source)

Experimental Setup

This study will leverage Amazon Web Services (AWS) cloud computing for the image classification, by loading the training and validation data to AWS S3 and use AWS EMR to create a PySpark cluster environment using GPU hardware for initial data exploration and training the networks. Pytorch will be primarily used for modeling the ConvNets. DCGAN implementation by Pytorch will be utilised for training the generator. For data visualization, Matplotlib and OpenCV2 will be used for plotting purposes.

Experimental Workflow

With respect to training the baseline model, the CheXPert training and validation images were resized to 224 x 224 pixels as the DenseNet121[3,6] model takes in the same size as it's input and an additional sigmoid layer was added as the last layer to output the probabilities of the 14 disease classes.The images were augmented using a series of transformations like RandomResizedCrop and RandomHorizontalFlip and normalized as per DenseNet standards, followed by conversion into a pytorch tensor object and fed to the model.A test set of 500 images were randomly sampled from the train set for model evaluation.

Binary Cross Entropy loss and Adam optmizer were chosen as the Loss function and Optimizer for the model. The learning rate for the loss function was set at 0.0001 with default β-parameters of $\beta 1$ = 0.9, $\beta 2$ = 0.999 as stated in Irvin, J. et al[3]. The batch size was set to 16.

The model was trained for 3 epochs and the model with the best average AUC scores for all 14 pathological cases on the validation set was saved for inference. The best model was then evaluated on the test set and the ROC curves for the 3 classes were plotted as shown in Figure 4 and Figure 5 for U-ones and U-zeroes approach respectively.

Training a GAN model is quite difficult as the minimax landscape is non convex and hence attaining convergence in practical is a daunting task. Another problem with GAN is called mode collapse, wherein the generator collapse and hence doesnot produce enough diverse samples. Mode collapse can be handled using multiple GANs. In this study, i choose to train ony single DCGAN model and train it for 1000 epochs to genrate 128 x 128 images. Though the Densenet121[3,6] model takes in 224 x 224 image as input, training a DCGAN to generate such high resolution images is quite difficult . Hence the 128 x 128 was resize to 224 x 224 for training the final classifier.

In this study, the GAN model were trained with Chest Xray images from 3 pathological classes, namely Pneumonia, Lung Lesion and Consolidation, as their percentage in the training dataset was low and hence causing imbalance. A batch size of 64 was chosen for the training. All model weights shall be randomly initialized from a Normal distribution with mean=0, stdev=0.02.

Figure 2 and 3 shows the basic architecture used in the construction of generator and discriminator. The generator maps the input latent space vector -z of size 100 to the training data space through a series of convolutional transpose layers, paired with a 2d batch norm layer and a relu activation. The generator output is fed to a tanh function to return it to input image range of [-1,1]. The discriminator uses a series of convolutional layers , batch norm layers and leaky relu activations and finally classifies the image as real or fake. Here, no max pooling layers are used as advised in[9] and instead strided convolutional layers was applied for the downsampling.

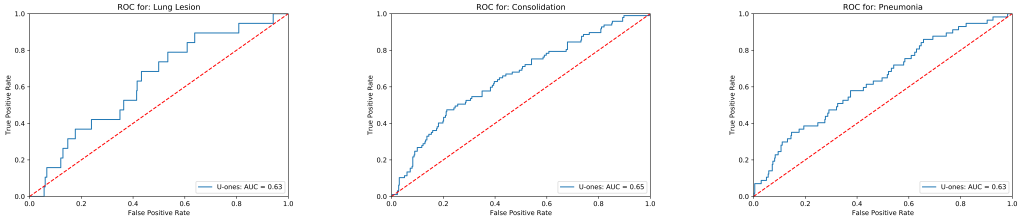
Both Generator and Discriminator uses Adam optimizer with a fixed learning rate of 0.0002 with $\beta 1$ as 0.5 and $\beta 2$ as 0.999.

After DCGAN training, the generator will be capable of generating artificial images for Pneumonia, Lung Lesion and Consolidation data space. Enough examples were generated so as to almost balance the distribution of these three pathological cases in the training dataset and the same was concatenated with the training set. The Final DenseNet121[3,6] classifier was kept same as the baseline model and was trained with the GAN enriched dataset and results were analysed.

Experimental Results and Evaluation

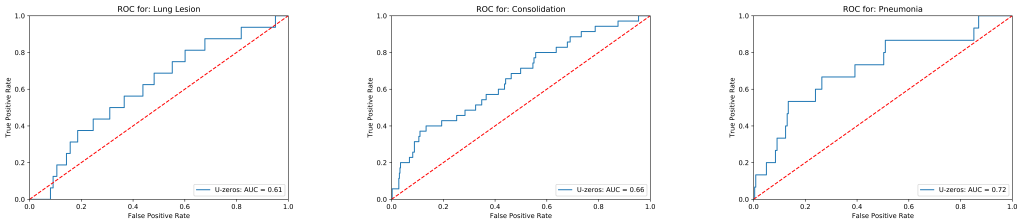
Baseline

As illustrated in Table 2, U-zeroes provided better ROC score for Enlarged Cardiomediastinum, Pneumonia and Pneumonothorax, while U-ones had better score for Pleural Other and both showed almost equal scores for other pathological classes. Hence, U-zeroes seems to be the better approach here. The highest ROC score of 0.78 was for No Finding class and the average score was around 0.65 in general. However, there is more room for improvement in the model training phase as the ChexPert baseline was able to achieve ROC scores greater than 0.81 for most of the pathological cases.



(a) ROC Curve for Lung Lesion **(b)** ROC Curve for Consolidation **(c)** ROC Curve for Pneumonia

Figure 4: Baseline ROC Curves for U-ones



(a) ROC Curve for Lung Lesion **(b)** ROC Curve for Consolidation **(c)** ROC Curve for Pneumonia

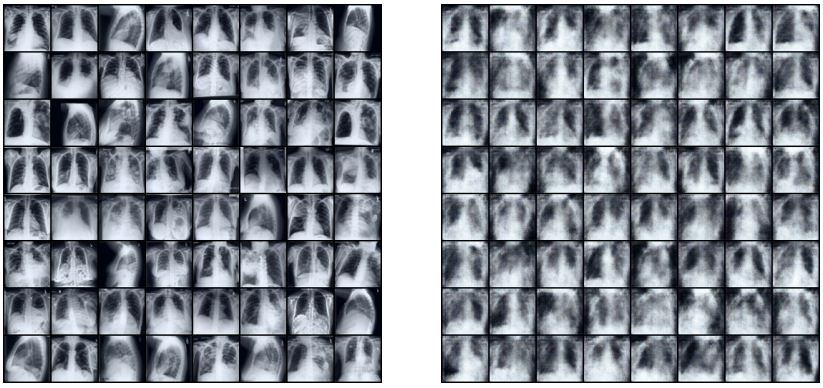
Figure 5: Baseline ROC Curves for U-zeroes

Pathology	**U-ones**	**U-zeroes**
No Finding	0.78	0.78
Enlarged Cardiomediastinum	0.51	0.62
Cardiomegaly	0.64	0.64
Lung Opacity	0.63	0.64
Lung Lesion	**0.63**	**0.61**
Edema	0.78	0.77
Consolidation	**0.65**	**0.66**
Pneumonia	**0.63**	**0.72**
Atelectasis	0.63	0.63
Pneumothorax	0.66	0.71
Pleural Effusion	0.67	0.66
Pleural Other	0.75	0.71
Fracture	0.67	0.67
Support Devices	0.74	0.74

Table 2: Average AUC ROC values for U-ones and U-zeroes approach

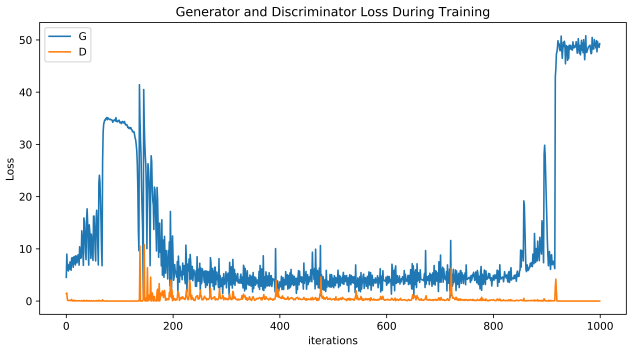
DCGAN

Figure 6 depicts the real training samples and fake images generated by the generator after around 400 epochs. Once the images generated looked near real, the training was commenced. Using the trained Generator, artificial images were generated from the Pneumonia, Lung Lesion and Consolidation data space and concatenated with the training dataset to make it nearly balanced for them. Afterwards, the same pretrained DenseNet121[3,6] classifier was trained on the GAN enriched training dataset and the results were analysed. Figure 7 helps to visualize the Generator and Discriminator loss progression. After around 900 epochs, the models lost stability as the generator started producing noisy images.



(a) Real samples from the training set **(b)** Fake Samples generated by the Generator

Figure 6: GAN - Real vs Fake Chest Xray Images

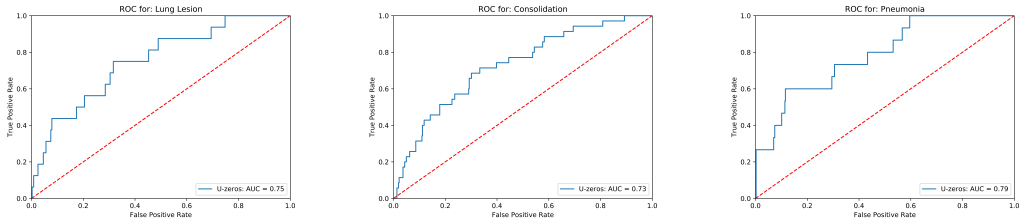


(a) Generator And Discriminator Loss Curve

Figure 7: Generator And Discriminator Loss Curve

Final Classifier

Figure 8 illustrates the final classifier ROC scores for Pneumonia, Consolidation and Lung Lesion for the u-zeroes approach. As evident from the plots and Table 3, an improvement in the ROC scores for the three pathological cases are observed.



(a) ROC Curve for Fracture **(b)** ROC Curve for Consolidation **(c)** ROC Curve for Pneumonia

Figure 8: Final Classifier ROC Curves for U-zeroes

Pathology	U-zeroes Baseline Classifier	U-zeroes Final Classifier
Lung Lesion	**0.61**	**0.75**
Consolidation	**0.66**	**0.73**
Pneumonia	**0.72**	**0.79**

Table 3: Final Classifier AUC ROC scores for U-zeroes approach

Conclusion and Challenges

In this project , i have attempted to set the baseline model stated in article Irvin, J. et al[3] using classic data augmentation techniques available inbuilt in Pytorch. As the baseline accuracy is lower than the ChexPert baseline, more data augmentation techniques might need to be applied here to achieve better ROC score. After setting the baseline, DCGAN implementation of Generative Adverserial Networks was applied to the Xray images of the imbalanced classes and the baseline model was trained again with the enhanced and balanced dataset generated using the Xray images of Pneumonia, Lung Lesion and Consolidation from GAN generator. The final classifier results showed that artificial data generated by GAN based unsupervised techniques can help enrich and balance real Medical imaging dataset, resulting in the substantial improvement in the Densenet121 classification performance of the 3 selected pathologial cases.

The biggest challenge faced was in terms of training the DCGAN model. GAN based training suffers from mode collapse, non-convergence , high sensitivity to hyperparameters and overfitting. Apart from these, it is also very challenging to generate high resolution GAN images and hence this study chose to generate 128 x 128 images, which were later rescaled to 224 x 224 for the final classifier training.

Inspite of the challenges, this study was able to suggest that data augmentation using aritificially generated images can help improve a classifier performance and help build better predictive models in the healthcare domain.

References

1. Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks." (2014). Web.
2. Salehinejad, H., Valaee, S., Dowdell, T., Colak, E.,& Barfett, J. (2017). Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks.
3. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., . . . Ng, A. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison.
4. Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." (2017). Web.
5. Wang, Xiaosong, et al. "ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." Vol. 2017, 2017, pp. 3462–3471.
6. Huang, G., Liu, Z., Van der Maaten, L., & Weinberger, K. (2016). Densely Connected Convolutional Networks.
7. Ausawalaithong, W., Marukatat, S., Thirach, A., & Wilaiprasitporn, T. (2018). Automatic Lung Cancer Prediction from Chest X-ray Images Using Deep Learning Approach.
8. Frid-Adar, et al. "GAN-Based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification." Neurocomputing, vol. 321, 2018, pp. 321–331.
9. Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.