CS230

Chest X-Rays Pathology Detection Using Augmented Datasets with GAN and Radiology Reports

Viveak Ravichandiran (SUNet ID: vravicha)
Aditya Srivastava(SUNet ID: adityaks)
Ying Chen (SUNet ID: smileyc)

Abstract

X-Rays are the most common and best available medical imaging technique used to diagnose lungs, heart and chest related diseases. The number of radiologists is decreasing in the U.S. [1], which is even worse in the underdeveloped countries. This motivated us to develop an AI solution by building a hybrid deep Neural Network architecture using augmented datasets with GAN and Radiology Reports to detect and recognize cardiopulmonary diseases to help radiologists to maximize their effort in diagnosing the problems.

1 Introduction

There are promising results of Chest X-Rays pathology detection and classification and here we try to improve it by building a hybrid deep Neural Network architecture using augmented datasets with GAN and Radiology Reports to detect and recognize cardiopulmonary diseases from both free-text radiology reports and Chest X-Rays to better understand different diseases. The motivation is to build an AI solution which would increasingly be accurate and help reduce the burden from the Radiologists by helping them prioritise their work, based on the results obtained by our solution. We hope that eventually we would have a solution which would be completely automated and surpass the performance of most of the Radiologists and be used in places where Radiologists are not available.

This is the first stage towards the final goal. To improve the prediction of datasets with imbalanced distribution, we presented a hybrid DNN model (GAN + ChexNet) to predict the probability of presence of each pathology by taking a chest X-ray image. The heatmap image that locates the area of diseases is also presented, as shown in Figure 1.

2 Related work

Past work have explored the potential solutions using deep learning. Wang. et al. (2017) [2], Yao. et al. (2018) [3] and Rajpurkar. et al. (2017) [4] developed deep learning CNN models (DenseNet-121) to classify different classes of diseases using chest x-ray images and achieving higher prediction precision than the previous work. Yan. et al. (2019) [5] exploited view-specific approach with two DenseNet-121 models on frontal and lateral views separately, which shows increased AUROC performance than previous DualNet model [6]. Irvin. et all (2019) [16] found DenseNet121 produced the best results compared to other CNN architectures for pathology detection. They also presented CheXpert dataset, in which the number of radiographs is about twice of NIH dataset.

Because of the good performance of DenseNet-121 model, we chose it as our baseline model. At the training stage, we also keep the class weight (determined by the sample counts of each class) in binary cross entropy, which is only used as pneumonia detection in CheXNet. At the testing stage,

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

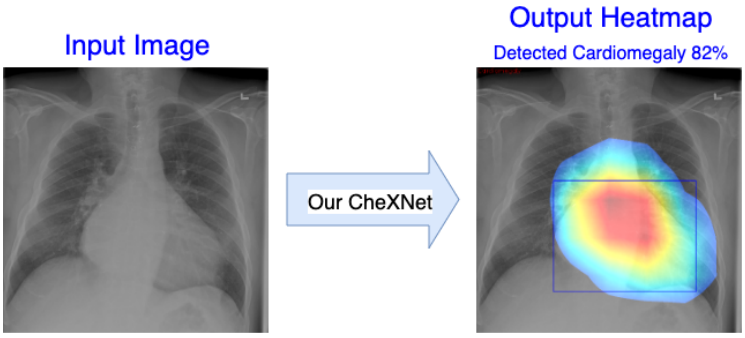


Figure 1: Diagram of input and output of our CheXNet model. This example shows correct detection of Cardiomegaly and heatmap importance matches ground truth location.

the end-to-end CheX-GAN and CheXNet models from radiology text report to pathology detection using MIMIC dataset, which has not been evaluated by other DL model.

Regarding data augmentation, Salehinejad. Et al. (2018) used GANs to generate effective artificial data [9] to improve prediction classification accuracy of imbalanced dataset [8] across five classes. Our GAN model is trained on the NIH dataset with all the 14 classes and generates synthetic images. However, compared to Salehinejad. Et al, our model feeds the generated images to the CheXnet model.

3 Dataset and Features

Two datasets were studied in this project - NIH and MIMIC-CXR datasets. NIH dataset was used at the beginning to test baseline model performance since this dataset has been used by several studies [2], [3], [4]. MIMIC-CXR is a new data set which was recently released by MIT which included both X-Ray images and free-text radiology reports.

3.1 NIH Dataset

The NIH dataset has 112k anonymized chest x-ray images of 30k patients from various age groups and genders across 18 disease categories including good tagging (i.e. No Finding) and some images have multiple tags. In this project, 14 classes were studied. Please refer to Table 3. for class names.

In the current NIH dataset, only frontal view images are included. Also the Data set has been split into Training/Dev/Test with 93:6:1 ratio.

Table 1: Data Split of NIH sets

Description/Data set	Training	Dev	Test
Number of images	104266	6336	1518
Split Percent of total	93	6	1
Good/Bad Ratio	53.90	53.35	51.25

3.2 MIMIC-CXR Dataset

The MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 is a large publicly available dataset of chest radiographs in DICOM format containing 377k images corresponding to 228k radiographic studies. The MIMIC-CXR consists of images (chest radio graphs with frontal or/and lateral view) and free-text reports. Since MIMIC-CXR dataset size is huge 4.7TB, we processed only a subset of x-ray images and converted it from DICOM format to png to feed our existing model. To test this dataset with CheXNet model, 7 different pathology classes mined from CheXpert NLP tool[14] is used to generate the labels - "Atelectasis", "Cardiomegaly", "Effusion", "Pneumonia", "Pneumothorax", "Consolidation", "Edema". Since images of this dataset have varied resolution and black order, the images were cropped at first to removed the black border dynamically and resized to 1024x1024, which is the same as image resolution of NIH dataset.

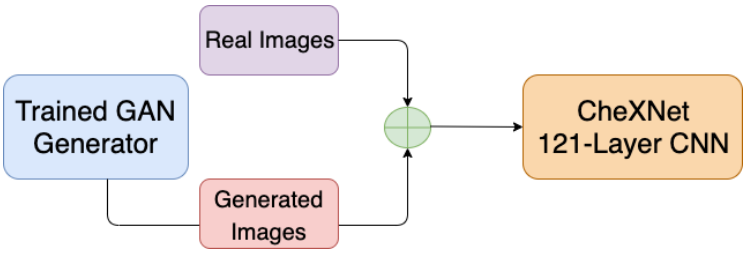


Figure 2: Architecture of CheX-Gan.

4 Methods

4.1 Baseline Model

Our baseline model is based on work of CheXNet algorithm [4]. CheXNet is a 121-layer Convolutional Neural Network (CNN)[11] that takes chest X-ray image as input, and outputs the probability of a chest pathology disease. CheXNet outputs a vector of binary labels indicating the absence or presence of each of the following 14 pathology classes: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, and Pneumothorax. The final fully connected layer in CheXNet with a fully connected layer produces a 14-dimensional output, after which an element-wise sigmoid non-linearity was applied. The final output is the predicted probability of the presence of each pathology class. The loss function to optimize the sum of weighted binary cross entropy losses where $p(Y_c = 1|X)$ is the predicted probability that the image contains the pathology c and $p(Y_c = 0|X)$is the predicted probability that the image does not contain the pathology c. $-w_+$ and $aligned - w_-$ represent the ratio of positive and negative cases in the training set, respectively.

$$L(X, y) = \sum_{c=1}^{14} [-w_+ \cdot -y_c \log p(Y_c = 1|X) - w_- \cdot (1 - y_c) \log p(Y_c = 0|X)]$$

The CheXNet model can be accessed here : `https://github.com/vivekravi/cs230-project/tree/master/CheXNet`. Please refer to config.ini file on the link for hyper parameters. This model is based on one existing Git repository[13], which is the our baseline model code for future improvement.

Medical image datasets are often highly imbalanced with over-representation of common medical problems and a paucity of data from rare conditions. The Chest X-Ray datasets used in this project also have imbalanced pathology classes.

4.2 GAN Model (CheX-GAN)

GAN models[15] are proven to improve performance of medical image classifications by simulating pathology images to overcome the class imbalance. Using GAN's, we are generating synthetic images based on the labeled dataset of Chest X-Rays to address this limitation. Our proposed CheX-GAN model generates artificial Chest X-ray images to balance pathology classes of NIH dataset and improve performance. The architecture of CheX-Gan is shown in Figure 2.

GANs are composed of two neural networks, a Generator G and a Discriminator D, which compete with each other over the available training data to improve their performance. The Discriminator network $D(x, \theta_d)$ receives a generated image $\hat{x}$ or a real chest X-ray x and produces an output $\hat{o}$, stating whether the input image is real or synthesized such that

$$\hat{y} = \frac{1}{1 + e^{-\hat{o}}} \qquad s.t. \hat{y} \in [0, 1]$$

where $\hat{y}$ 0 and $\hat{y}$ 1 state that the input chest X-ray is synthesized or real, respectively. The Generator network G trains so as to propose artificial images that the Discriminator network D(x) cannot distinguish from real images. The adversarial competition between G and D can be represented as

$$\min_G \max_D \mathcal{L}(D, G) = \mathbf{x} \sim pata\mathbb{E}(\mathbf{x})[\log D(\mathbf{x})] + \mathbf{z} \sim \mathbf{p_z}(\mathbf{z})\mathbb{E}[\log(1 - D(G(\mathbf{z})))]$$

5 Experiments/Results/Discussion

For training, The initial learning rate is 0.001. If the validation loss doesn't decrease for one epoch, learning rate will be reduced by a factor of 10. The Adam optimizer and default exponential decay rate parameters are used - 0.9 for the 1st moment estimates and 0.999 for the 2nd moment estimates.

5.1 Model Evaluation - AUROC

The AUROC values of current DCNN model is shown in the Table 2. This project started from pre-trained weights. However, AUROC values were very low, (as low as 0.5) for some classes, which was not consistent with the author's description. The pre-trained weights were used as the staring point for training dataset to re-train the model and update weights. After that, the AUROC scores improved overall.

Compared with original CheXNet results, the overall AUROC scores are slightly lower. This could be due to different test datasets. Original CheXNet has 420 images in test set. We used 9568 images in the test set. Compared to NIH dataset, the test results of MIMIC datasets show overall similar performance on Atelectasis, Cardiomegaly and Effusion classes, but worse on Pneumonia, Pneumothorax, Consolidation and Edema. There could be some distribution difference on these classes between two datasets. The lowest score of Pneumonia could be due to often vague appearance of pneumonia in X-ray images and overlap with other diagnoses[4].

Table 2: Results - AUROC

Pathology	**CheXNet (Rajpurkar et al. 2017)**	**Pre-trained Weights**	**CheXNet (ours) - NIH dataset**	**CheXNet (ours) - MIMIC dataset**	**CheX-GAN (ours) - Synthetic**
Atelectasis	0.809	0.821	0.784	0.834	-
Cardiomegaly	0.925	0.500	0.890	0.824	-
Effusion	0.864	0.889	0.852	0.883	-
Infiltration	0.735	0.722	0.713	-	-
Mass	0.868	0.841	0.844	-	-
Nodule	0.780	0.744	0.780	-	-
Pneumonia	0.768	0.661	0.657	0.594	-
Pneumothorax	0.888	0.884	0.767	0.621	-
Consolidation	0.790	0.747	0.797	0.749	-
Edema	0.888	0.668	0.865	0.760	-
Emphysema	0.938	0.578	0.974	-	-
Fibrosis	0.805	0.543	0.724	-	0.623
Pleural Thickening	0.806	0.792	0.739	-	0.564
Hernia	0.916	0.500	0.733	-	-

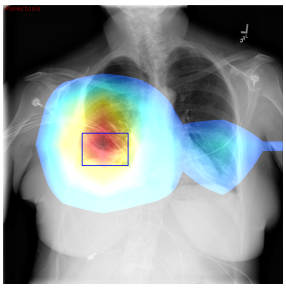
5.2 Model Interpretation

To interpret the model predictions and further evaluate how it correlates with the location of the pathology, the heatmap plot is also generated using class activation mappings [4]. An image is used as the input of the trained model, and then feature map is extracted from the batch normalization layer of the last convolutional layer. The following equation is used to compute class activation mapping.

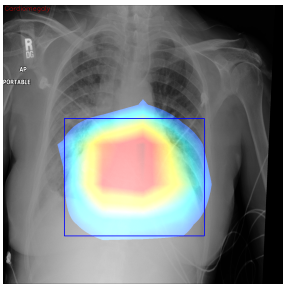
$$ActMap = \sum_{i=1}^{1024} w_i f_i \qquad w_i \in w_{i,c}$$

where $w_{i,c}$ is the weight for the ith feature map $aligned f_i$ in the last classification layer belonging to the pathology class c.

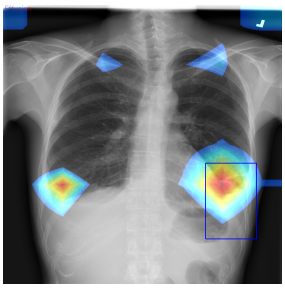
Thus, the map of most important features in the model prediction is obtained. Finally, the heatmap plot can be generated by overlaying normalized scaled $ActMap$ and original input image. In our experiment, the array shape of $ActMap$ is (7, 7) and image resolution is (1024, 1024). Several of example heatmap images are shown in Figure 3.



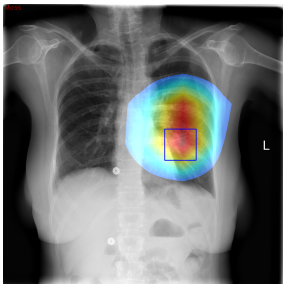
(a) Atelectasis



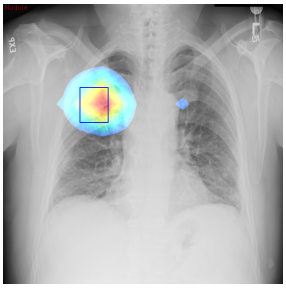
(b) Cardiomegaly



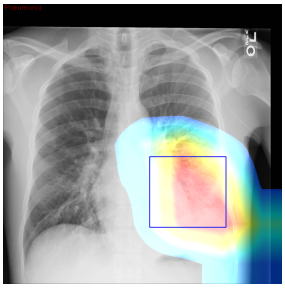
(c) Effusion



(d) Mass



(e) Nodule



(f) Pneumonia

Figure 3: Examples of heatmap images of different pathologies. The bounding box is ground truth region provided.

5.3 Limitations

Due to storage and compute limitations, we were not able to process the entire MIMIC dataset because of high resolution DICOM images which required more computational resource to downsample and feed to the model for training.

The model prediction accuracy could be improved by incorporating image training of other view positions. We used only frontal view images from NIH datasets for training. The model performance could be improved for MIMIC dataset with adding it into training stage.

6 Conclusion/Future Work

By addressing the class imbalance problem in medical imaging or chest X-ray in general, using similar datasets and generating augmented/synthetic images using GAN, helped to improve the performance of some of the pathology classes.

If we continued to work on this project for next 6 months, we would process more images and radiology reports from the new MIMIC-CXR dataset to improve the predications of all the classes. By processing the radiology reports, the model can also be trained to predict secondary diseases.

7 Contributions

All team members worked together and contribute equal amount of efforts to the project. Ying Chen mainly worked on baseline ChexNet and NLP model study, processing MIMIC datasets, and model evaluation. Aditya mainly worked on processing NIH dataset, design discussions. Viveak mainly studied the MIMIC CXR dataset and helped to get access, setup the AWS instance, build a GAN model to generate synthetic x-ray images for data augmentation. We all collaborated on building Interim and Final reports as well as the poster.

References

[1] Douglass Margaret, *Computer-assisted de-identification of free-text nursing notes*. Master's Thesis, 2005. MIT.

[2] Wang Xiaosong, Peng Yifan, Lu Le, Lu Zhiyong, Bagheri Mohammadhadi and Summers Ronald M., *Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases*. arXiv preprint arXiv:1705.02315, 2017.

[3] Yao Li, Poblenz Eric, Dagunts Dmitry, Covington Ben, Bernard Devon and Lyman Kevin., *Learning to diagnose from scratch by exploiting dependencies among labels*. arXiv preprint arXiv:1710.10501, 2017.

[4] Rajpurkar Pranav, Irvin Jeremy, Zhu Kaylie, Yang Brandon, Mehta Hershel, Duan Tony, Ding Daisy, Bagul Aarti, Ball Robyn, Langlotz Curtis, Shpanskaya Katie, Lungren Matthew and Ng Andrew. *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. arXiv:1711.05225v3, 25 Dec 2017.

[5] Yan Michael, Chang Ying and Ang Yu. *CheXDualNet: A View-Specific Approach to Chest Pathology Classification*. Stanford CS230 project, Spring 2019.

[6] Rubin Jonathan , Sanghavi Deepan, Zhao Claire, Lee Kathy, Qadir Ashequl, Xu Minnan. *Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks*. arXiv:1804.07839v2 [cs.CV] 24 Apr 2018.

[7] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, Maria de la Iglesia-Vayá. *PadChest: A large chest x-ray image dataset with multi-label annotated reports*.

[8] Salehinejad Hojjat, Valaee Shahrokh, Dowdell Tim, Colak Errol and Barfett Joseph. *Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks*. arXiv:1712.01636v2 [cs.CV] 12 Feb 2018.

[9] Tang Weixuan, Tan Shunquan, Li Bin and Huang Jiwu. *Automatic steganographic distortion learning using a generative adversarial network*. IEEE Signal Processing Letters, vol. 24, no. 10, pp. 1547–1551, 2017.

[10] He Kaiming , Zhang Xiangyu, Ren Shaoqing and Sun Jian. *Deep Residual Learning for Image Recognition*. arXiv:1512.03385v1 [cs.CV] 10 Dec 2015.

[11] Huang Gao, Liu Zhuang, Maaten Laurens, Weinberger Kilian. *Densely Connected Convolutional Networks*. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700-4708.

[12] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. *Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment*. arXiv:1907.09085v2 [eess.IV] 23 Jul 2019.

[13] Bruce Chou. *GitHub Repository: CheXNet-Keras*. `https://github.com/brucechou1983/CheXNet-Keras`

[14] Irvin Jeremy, Rajpurkar Pranav, Ko Michael, Yu Yifan, Ciurea-Ilcus Silviana, Chute Chris, Marklund Henrik, Haghgoo Behzad, Ball Robyn, Shpanskaya Katie, Seekins Jayne, Mong David, Halabi Safwan, Sandberg Jesse, Jones Ricky, Larson David, Langlotz Curtis, Patel Bhavik, Lungren Matthew and Ng Andrew. *GitHub Repository: chexpert-labeler*. `https://github.com/stanfordmlgroup/chexpert-labeler`

[15] 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*. arXiv:1712.01636v2 [cs.CV] 12 Feb 2018.

[16] Irvin Jeremy, Rajpurkar Pranav, Ko Michael, Yu Yifan, Ciurea-Ilcus Silviana, Chute Chris, Marklund Henrik, Haghgoo Behzad, Ball Robyn, Shpanskaya Katie, Seekins Jayne, Mong David, Halabi Safwan, Sandberg Jesse, Jones Ricky, Larson David, Langlotz Curtis, Patel Bhavik, Lungren Matthew and Ng Andrew. *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison.* arXiv:1901.07031 [cs.CV] 21 Jan 2019.