Detecting pathology from Chest X-Ray Images

Das, Abhishek K
Computer Science Department
Texas A&M University
College Station, Texas
abkds@tamu.edu

***Abstract*—In the past few years, deep learning methods have been driven by large labeled datasets. Large datasets helped in creating systems which reach expert-level efficiency. Stanford recently published the chexpert dataset which contains 224,316 chest radiographs of 65,240 patients. The dataset was labeled by a labeler which captured various different pathology present in each x-ray image. A validation set of 200 images has been provided which was labeled by 3 radiologists. The ROC curves for the performance of the network have been released in this paper.**

***Index Terms*—deep-learning, dense-net, cnn**

I. Introduction

Chest radiography is the most common imaging method globally for, critical for screening and diagnosis of many diseases. Automated interpretation of the chest radiographs can immensely improve the workflows of healthcare systems and help in large scale implementation of such systems. A large dataset for the same was needed.

The chexpert dataset solves this problem by providing 224,316 chest radiographs of 65,240 patients. The dataset has been labeled for 14 most common chest radiograph observations. All the radiographs originally has reports from which the labels were extracted using an automated labeler, an uncertainty was assigned to them depending on the label existing in the report. Following is the distribution of labels among the various pathology.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

TABLE I: Distribution of data

II. Label Extraction and Radiology Report

Stanford used a automated rule based labeler that scans the reports for pathology labeling and does that in 3 stages: mention extraction, mention classification and mention aggregation. On almost all the reports this method works better than the NIH [1] labeler having a higher F1 score.

(a) Frontal (b) Lateral

Fig. 1: The task is to predict the probability of the various pathology from the frontal and lateral chest x-ray images

III. Model

Uncertainty Approaches

The training labels in the dataset contain the labels 0 (negative), 1 (positive) and u (uncertain). We explore various methods to deal with uncertainty so that we can train our models.

Ignore In the ignore method we simply ignore the uncertainty labels in the dataset and only consider the positive and the negative points in the dataset but this effectively reduces the dataset by a huge margin in all the pathology. Take for example Cardiomegaly, if we ignore the uncertain labels we are left with positive 12.26% and negative 3.52%. While using this approach the loss function which we will be using is effectively

$$L(X, y) = \sum_o 1\{y_o \neq u\}[y_o \log p(Y_o = 1|X) + (1 - y_o) \log p(Y_o = 0|X)] \quad (1)$$

In the equation X is the input image and y_o is the output label which we are trying to predict. The indicator function is an indicator that we are ignoring the labels which are uncertain.

Binary Mapping As the name suggests, either we consider all the uncertain labels as (0) negative, the zeros-model or we

can consider all the uncertain labels as positive (1) the ones-model. These approaches are similar to imputation strategies (Kolesov et al. 2014) [2]. This approach can possibly alter the decision making capability of the system if the uncertain images contained semantically useful information.

3-Class classification In this approach, we consider each of the 3 classes separately rather than mapping the uncertainty label to either positive or negative. We consider the probability of each class $\{p_0, p_1, p_u\} \in [0, 1]$ and $p_0 + p_1 + p_u = 1$. We set up the loss as the categorical cross entropy over the classes and train the network. At test time, we output the probability of the positive label by ignoring the output of the uncertainty label and taking the softmax over the remaining two.

Augmentation Since we suffer from unbalanced data in the dataset, to balance we try the oversampling approach [4]. Using augmentative methods [3] we oversample the underrepresented by creating new images. Depending on the number of the classes we consider (2 or 3) we try to balance the dataset accordingly. We then train the model on this augmented dataset. Augmentation methods are common in the image domain because the augmented images are either rotated or flipped which doesn't matter in our case also, as the x-ray image can be flipped/rotated and that doesn't affect the pathology in the given x-ray.

IV. Training

For the various uncertainties, the same architecture and training process was used. Experiments were carried out using various convolutional neural network architectures like ResNet152, DenseNet121, DenseNet169, and Inception-v4. It was found that DenseNet121 and DenseNet169 performed better than other architectures. Images were feed into the network after resizing as 320×320 pixels. In the case of augmentation of the images, were zoomed to a scale ranging from 1.05 to 1.2 with a probability of 0.3, subsequently rotated to $\pm 25°$ with a probability of 0.7 and then flipped alongside the horizontal axis with a probability of 0.7. For optimization, the Adam optimizer was used with the default β parameters, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 1×10^{-4}. At each epoch the dataset was shuffles and batches were sampled with a size of 16. The training was done for 10 epochs (at max). Callbacks for stopping early, when there is no improvement were given to the optimizer. Training was done on Tesla V100 GPU and each training took roughly 7 hours to complete.

V. Validation Results

Validation Set

The validation set contains a total of 200 studies from 200 patients. These images were randomly sampled from the dataset. Three board-certified radiologists individually annotated each image in the validation set, marking whether a pathology is present or absent or unlikely. These annotations were then binarized so that the uncertain and positive cases are treated as positive and the negative cases as negative. The majority among these binarized annotations is then used as the final ground truth.

A. Figures

We take two of the most important pathology and plot the receiving operator characteristic area under the curve for both of them under different methodologies. For the cases of atelectasis, the area under receiving operator characteristic curve increases significantly from 0.787 to 0.828. The following plot depicts the same.

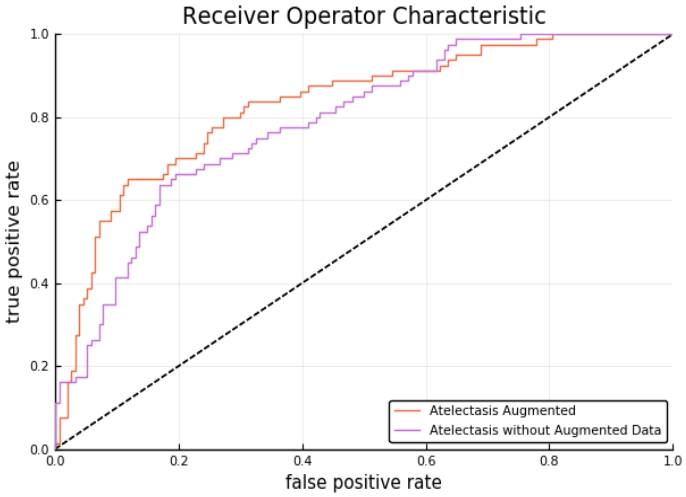

Fig. 2: Receiving Operator Characteristic Curve for Atelectasis

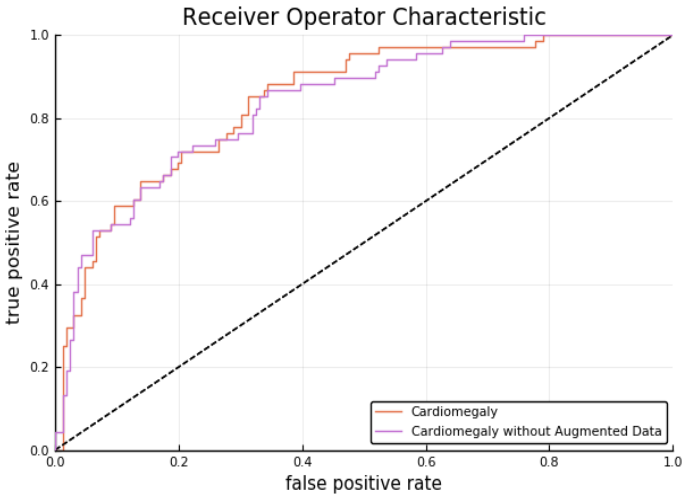

Fig. 3: Receiving Operator Characteristic Curve for Cardiomegaly

Now for the case of cardiomegaly we have similar improvements in the receiving operator characteristic area under the curve. The area under the curve for non augmented data is 0.837 where as the for augmented date it is 0.844. We see that balancing the dataset increases the ROC-AUC score

as compared to the unbalanced dataset. Above curve depicts the exact ROC curve for cardiomegaly.

The table below compiles the ROC-AUC scores for the experiments carried out.

	Cardiomegaly	Atelectasis
Non-Aug	0.837	0.787
Augmented	0.844	0.828

VI. Future Work

Images also contain global level interactions which apart from the regional similarities caputured by convolutional filters can be collected by fully connected layer, but fully connected layers introduce a lot of parameters and generally grossly overfit and worsen the network's performance. In recent developments, Attention (Vaswani et al) [1] has shown promising results when applied on images to capture global interactions without overfitting and including much lesser parameters. In our next set of experiments, attention will be used to try to improve the network performance.

VII. Links and Resources

Here is the link for github repository.

All the experiments were done on floyd hub, it will take too much if run locally without a GPU. Following is the command used for running the atelectasis pathology, similar files for other pathology are there.

```
floyd run --gpu2 --env keras --data abkds/
    ↪ datasets/atelectasis/1:atelectasis
    ↪ --data abkds/datasets/
    ↪ chexpert_validation/1:
    ↪ chexpert_validation 'python␣
    ↪ atelectasis.py'
```

To run the gui demo open the folder and type in terminal

```
python3 gui_chexpert_rl_python3.py
```

Here is the link for the demo

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin "Attention Is All You Need"

[2] Kolesov, A.; Kamyshenkov, D.; Litovchenko, M.; Smekalova, E.; Golovizin, A.; and Zhavoronkov, A. 2014. "On multilabel classification methods of incompletely labeled biomedical text data. Computational and mathematical methods in medicine 2014."

[3] Marcus D. Bloice, Christof Stocker, and Andreas Holzinger, "Augmentor: An Image Augmentation Library for Machine Learning,"

[4] Mateusz Buda, Atsuto Maki, Maciej A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks"