Lung Disease Detection and Classification with 3D Convolutional Neural Network

M.H. Motamedi
motamedi@rowan.edu
Rowan University, College of Engineering
Glassboro, NJ, USA

H. Alfergani
alferganh3@students.rowan.edu
Rowan University, College of Engineering
Glassboro, NJ, USA

Abstract:

This manuscript demonstrates a computer-aided diagnosis (CAD) system for a ChestX-ray dataset comprises 112,120 frontal-view X-ray images of 30,805 unique patients with the text-mined fourteen disease image labels. The initial approach was to directly feed the downsized X-ray images (28X28) into 3D Convolutional Neural Networks (CNNs) using two-layer inception model for multi-label classification, but this proved to be inadequate and result shows a significant overfitting. Instead, a modified binary dataset (disease and no-disease), allowing more efficient training and detection and more generalizability to other Lung diseases, was used to detection of disease from resized images. The 3D CNNs produced a total test set accuracy of 78% which includes 87% and 68% accuracy of true negative and true positive, respectively. Although the model is pretty light that can be performed even with multiple CPUs processor, the accuracy of results is acceptable, knowing that the false negative error is achieved by less than 13%.

1. Introduction:

1.1 Background:

General thoracic diseases are diseases that occur in your lungs, esophagus, or chest. Typically, these diseases are treated costly by cardiothoracic surgeons who specialize in general thoracic surgeries. Notably, over 12 million people in the UK have been diagnosed with a lung condition, and lung disease is the nation's third biggest killer. In the U.S., lung cancer is one of the most common cancers, accounting for over 225,000 cases, 150,000 deaths, and $12 billion in health care costs yearly [1].

Our task is a binary classification problem to detect the presence of thorax disease in patient ChestX-ray images with and without thorax disease. The rapid and tremendous progress has been evidenced in a range of computer vision problems via deep learning and large-scale annotated image datasets [2, 3, 4, 5]. Drastically improved quantitative performances in object recognition, detection and segmentation are demonstrated in comparison to previous shallow methodologies built upon hand-crafted image features. We aim to use methods from computer vision and deep learning, particularly 2D and 3D convolutional neural networks, to build a computationally light classifier with an acceptable predictive capacity. An accurate Thorax disease classifier could speed up and reduce costs of disease screening, allowing for more widespread early detection and improved survival. In this study, the computer-aided diagnosis (CAD) system is constructed such that it takes patient chest X-ray images as input and outputs is whether or not the patient has thoracic diseases.

1.2 Related Work:

There have been recent efforts on creating openly available annotated medical image databases [6, 7, 8, 9] with the studied patient numbers ranging from a few hundreds to two thousands. Particularly for chest X-rays, the largest public dataset is OpenI [10] that contains 3,955 radiology reports from the Indiana

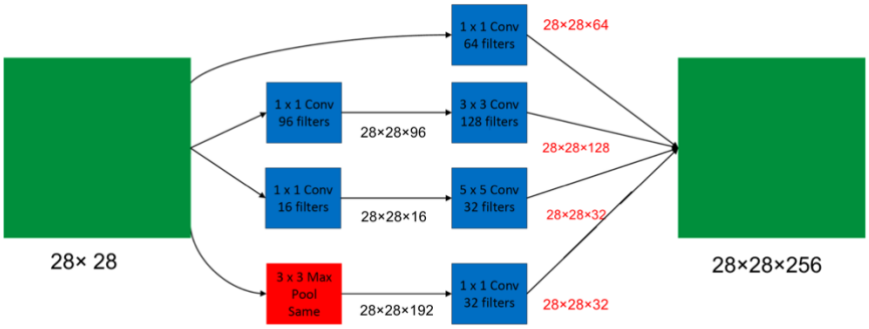
Network for Patient Care and 7,470 associated chest x-rays from the hospitals picture archiving and communication system (PACS). This database is utilized in [11] as a problem of caption generation but no quantitative disease detection results are reported. This newly proposed chest X-ray database is at least one order of magnitude larger than OpenI [10].

2. Methods:

A typical Convolutional neural network (CNN) is made up of stacked convolutional layers in combination with max pooling and dropout. For larger datasets such as Imagenet, deeper architectures are used to get better results and dropout is used to prevent overfitting. In a typical CNN layer, we make a choice to either have a stack of 3x3 filters, or a stack of 5x5 filters or a max pooling layer. In general all of these are beneficial to the modelling power of the network. The inception module suggests the use of all of them. This means instead of adding a particular filter size layer, we add all 1x1, 3x3, 5x5 filters and perform convolution on the output from the previous layers. Since pooling has been essential for the success of current CNNs, the inception module also includes an additional pooling path. The output of all the filters are concatenated and passed on as input to the next layer.

2.1 Architecture

In this study, we used a two-layer inception model that contains in total 702,386 trainable parameters. The two layers have the same structure and they are connected together sequentially. Below figure represents the architecture of each inception layer.



A layer of Inception module

Notice that the model includes the variety of convolutions that can capture different unique features within an image; specifically, we used 1×1, 3×3, and 5×5 convolutions along with a 3×3 max pooling layer. The larger convolutions are more computationally expensive, so here first we applied a 1×1 convolution reducing the dimensionality of its feature map, then we passed the resulting feature map through a Relu function and at the end, we made the larger convolution (in this case, 5×5 or 3×3).

3. Results:

3.1. Part I: 14 Lung disease classification

Our first goal was to classify 14 different lung diseases. However due to computation power limitations and time need to process and re-adjust each hyper parameter, we were not able to get a good test accuracy. The models (ResNet 20, ResNet44 and two inception) that we have used showed overfitting where the test accuracy kept bouncing between 50-60% and the training accuracy above 95% most of the times. Classification and labels with the number of images for each class are show in Table 1, where last Label (15) collecting all images that have two or more combined lung diseases.

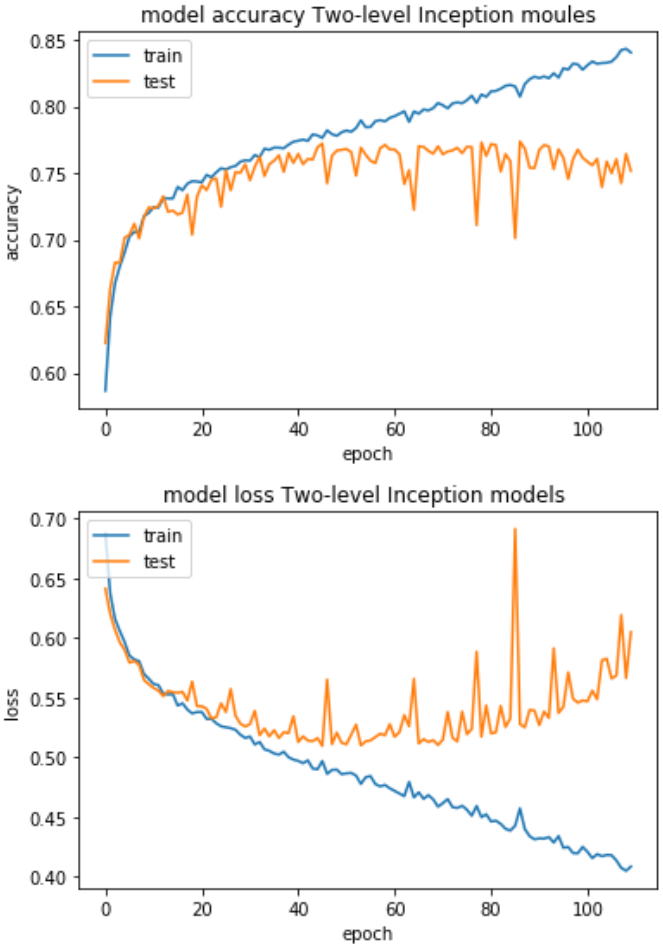
Diseases	Label	No of images
Atelectasis	0	4215
Cardiomegaly	1	1093
Consolidation	2	1310
Edema	3	628
Effusion	4	3955
Emphysema	5	892
Fibrosis	6	727
Hernia	7	110
Infiltration	8	9547
Mass	9	2139
No Finding	10	60360
Nodule	11	2705
Pleural_Thickening	12	1126
Pneumonia	13	322
Pneumothorax	14	2194
combine 2 or more diseases	***15***	20795
Total no. of images		112118

3.2. Pat II: Binary classification

In this part we used two inception models for binary classification of the given data. That is to classify the image as either there is a Lung Diseases or no lung disease. For the binary classification of Lung Disease or no Lung disease we reclassified the images in Table 1 as shown in Table 2, where all No finding class in table1 is classified as no disease and the rest of the image is classified as Lung disease with data split between training and test 70:30 respectively.

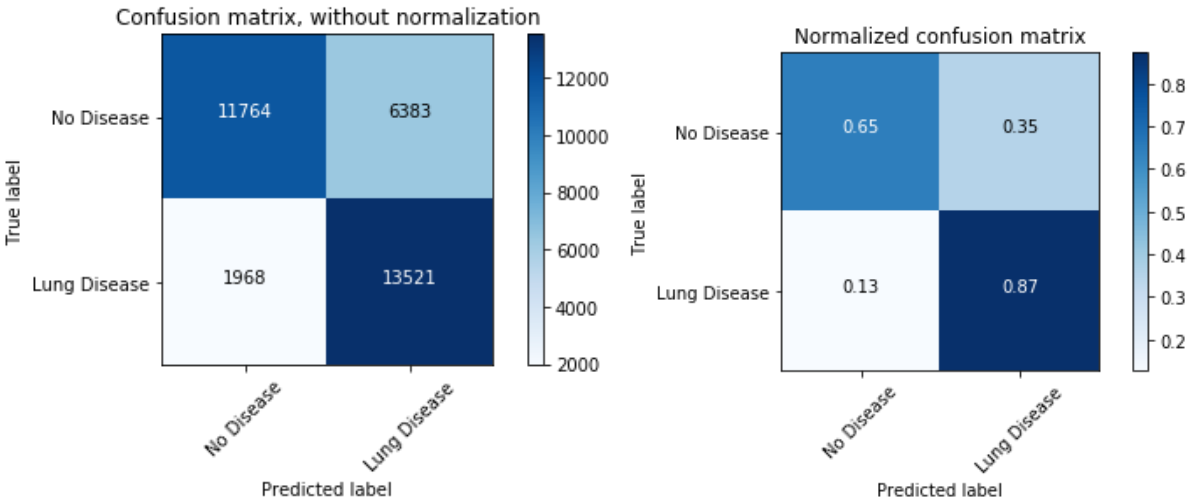
Class	No of labels
No disease	60360
Disease	51758
Total	112118

The results obtained from the binary classification show that the test accuracy is just above 75% and loss around about .55



Model performance on training and testing datasets: accuracy and loss

To investigate the problem further we have calculated the confusion matrix as follows



The diagonal elements in this matrix shows that out of 33636 test data (30% out of the all data), No-diseases class (i.e. true negative TN) is classified correctly with 65% accuracy and the Lung disease

class (i.e. true positive TP) is classified correctly with 87%. These give us the total test accuracy of 75.17%. However, off diagonal elements shows the False negative (FN) with 13% of the data and False positive (FP) classified is 35%.
The results of binary classification for each class can be summarized in the following Table 3

Class		precision	recall	f1-score	support
No Disease	**0**	0.87	0.65	0.74	18147.00
Lung Disease	**1**	0.65	0.87	0.76	15489.00
Avg / Total		0.78	0.75	0.75	33636.00

Where:

Precision: is the ratio of correctly predicted positive observations to the total predicted positive observations and can be calculated by

Precision = TP/TP+FP or TN/TN+TP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class –

Recall = TP/TP+FN

F1 Score is the weighted average of Precision and Recall.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

The recall values for both classes are both above .5 which considered a good a value. The F-score results is more indicative here than the accuracy since the FN and FP *(the off-diagonal elements) are not evenly distributed (not equal)

4. Discussion:

The X-Ray- 14 dataset is a hug dataset with 112120 images of size 1024x1024, it needs a high-performance computer, many GPU's and large memory to access it. For image size 28x28 the dataset size is approximately 60 MB. With image size 128x128, the dataset is about 6GB and it is about 55 GB for 1024x1024. The founders of this dataset have labeled the data based on NLP algorithms and they claimed about 90% labeling accuracy. To obtain a better accuracy, images need to be visually checked and discard images with low resolution or un readable images and to adjust some labels where NLP does not correctly label the images with the right disease. The available resources for us can handle only 28x28 image size which we do not think it will give satisfactory results as we are losing some features by resizing the image from 1024x 1024 (original dataset) to 28x28, given that the original image was 3000,2000 pixels. Another way to improve it we need to use data augmentation as some labels has small number of images as shown in Table -1 , some have few hundreds and most of them have less than 5000 image per label, which is not enough to train the network to give reasonable prediction accuracy.

The available resources enable us to perform binary classification by grouping the labels in two categories disease or no Lung disease. The results were satisfactory up to a certain extent and still can be improved if we increase the levels of inception models and changing the hyperparameters.

5. Conclusion:

The X-Ray-14 dataset is very challenging in building CNN with better accuracy in terms of image size and number of images. Despite the hug number of images (112120 image0 but there is still some labels has insufficient images to get a reasonable results. The Authors and the founders of this dataset have shown accuracy of 70 – 85% for some classes for only eight Lung disease. The conclusion is that there is still a lot to be done to enhance the accuracy of this dataset.

References:

[1]- Centers for Disease Control and Prevention, "Lung cancer statistics." https://www.cdc.gov/cancer/lung/statistics/, 2016.

[2]- A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[3]- O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, pages 115(3): 211–252, 2015.

[4]- M. Everingham, S. M. A. Eslami, L. J. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, pages 111(1): 98–136, 2015

[5]- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and L. Zitnick. Microsoft coco: Common objects in context. ECCV, pages (5): 740–755, 2014

[6]- H.-J. Wilke, M. Kmin, and J. Urban. Genodisc dataset: The benefits of multi-disciplinary research on intervertebral disc degeneration. In European Spine Journal, 2016

[7]- J. Yao and et al. A multi-center milestone study of clinical vertebral ct segmentation. In Computerized Medical Imaging and Graphics, pages 49(4): 16–28, 2016.

[8]- H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In MICCAI, pages 520–527. Springer, 2014.

[9]- H. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In MICCAI, pages 556–564. Springer, 2015.

[10]- Open-i: An open access biomedical search engine. https: //openi.nlm.nih.gov

[11]- H. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In CVPR, 2016.

Github-link:

https://github.com/Alfergani1965/DL-Class

https://github.com/mhmotamedi/LungDiseaseClassifier