ChestX-Ray8 Results

1 Introduction

The X-ray database "ChestX-Ray8" was introduced by [Wang et al., 2017] recently as a new dataset for study through the National Institute of Health. This databse contains 108,948 front X-ray images of 32,717 patients. Each of these images are labeled with a series of 15 different diagnoses, including no finding. The designations are Cardiomegaly, Emphysema, Effusion, Hernia, Infiltration, Mass, Nodule, Atelectasis, Pneumothorax, Pleural Thickening, Pneumonia, Fibrosis, Edema, Consolidation, and No Finding.

The purpose of this self study is to attempt to use supervised learning to fit a neural net, specifically with Tensor Flow/Keras, to assign a prediction "likelihood" to the presence of each disease in these x-ray images. Further we enlist this model to create a heatmap of pixels in the images that most affected the decision to label each image as such, for example if the model designates that a hernia is likely present the heatmap image will display which pixels were instrumental in coming to this conclusion.

This is being written with the intention of being intelligible to relatively laypeople, with an emphasis on those not familiar with statistics, or machine learning. A more technical description is available in the paper cited above.

2 A Brief Overview

A brief explanation of what the model is doing, a simple explanation of how it works, and whats to be expected is warranted. We are using images of x-rays to attempt to train a neural net to predict the presence of different illnesses. Naturally in reality the presence of an illness is a binary term, it is either present or it isn't. In statistics this level of certainty is impossible, however we can train the model to give us an inclination to how strongly it feels the illness is present. For example, if we enter an X-ray into the model and it says it thinks there is a 20% chance of edema, we can ignore it, but if it says there is 90% chance of pneumonia it is worth further scrutiny. Ultimately this is the goal, a tool to double check the work of doctors, and refer patients for further review.

The model does this by incorporating a 224 x 224 pixel image, taking each of these pixels (which are in black and white) and assigning them a number between 0 and 1, 0 if it is completely white and 1 if it is completely black.

The neural net then runs this grid through several training layers to find what combination of shades correlates to diseases.

We take this one step further, by also using a Gradient-weighted Class Activation Mapping (Grad-CAM) [R. Selvaraju and Batra, 2019] to create a heatmap of these diagnoses. The idea is this heatmap will point out which pixels motivated it to make the conclusion that it did. These heatmaps can be used to point out to medical professionals specific areas of interest when they are diagnosing a patient.

3 Methodology

3.1 Model Fitting

The implementation of the aforementioned neural network is fairy straight forward, although we must make a few preparations beforehand. Firstly, the original images are of the resolution 1024x1024. This is a relatively high resolution image which unfortunately is simply too dense to run on a personal computer. To alleviate the computational complexity we are forced to downscale the images to 224x224 resolution. Naturally this will hamper results, but will make the process feasible. Secondly, the data is labeled jointly, with each image having a list of ailments affecting the patient. To enlist the model we have to do some data cleaning, and reorganizing, essentially dividing the data into a table where each column is a diagnosis, and a 1 or 0 is present for whether or not it is present.

With this in order, we divide the data into a training, validation, and test sets which each consist of 70%, 15% and 15% of the data respectively. The idea behind this split is to train the data on the largest portion, use the validation data to assist in training, and finally to use the remaining set to test the efficacy of the model. This process, of dividing the data into a test and training set, is typical in statistics to prevent over fitting, and promote generalization, since the model never sees the test data until it is evaluated it acts as a trial run of how the model should behave in a future setting.

3.2 Heatmap Generation

Neural networks famously have a lack of interpretability. That is, they tend to give you a result but do not always present why that result occurred, or even how the model came to that result. With image data we are able to utilize Gradient-weighted Class Activation Mapping (Grad-CAM) [R. Selvaraju and Batra, 2019] to yield some insight as to what lead the model to come to its conclusion.

A simple understanding of this process is we look at a specific layer in our model, you can think of this as a step. In this step, we see how much our error changes for each covariate, in this case each covariate is one pixel. This process changes the conclusion for all classes, in this case each diagnosis, so we project this onto the diagnosis we are studying. This tells us which pixels are causing

the greatest decrease in error, which can be used to create a heat map that gives us a visual representation of the models process.

In our case, because we are using images that are 224 x 224, and our neural network has 5 convolution layers, the resulting heat map is 7 x 7. This is relatively sparse for the image, we apply some smoothing to help but will be noticeable on the images. If we were able to use the original image size of 1024 x 1024 the resulting heatmap would be 32 x 32 which would be considerably finer. This annoyance is somewhat alleviated by using smoothing packages on our 7 x 7 heatmap, but is simply a limitation of the computation power available.

4 Results

The result of the model fitting can be seen in Figure 1 and Table 1, which shows the ROC curve for each class, and AUC for each class respectively.

ROC and AUC curves are validation tools for binary classification, that is any data where you are attempting to predict whether something is present or not. In this form of classification, there is always a trade off between the true positive rate (Sensitivity) and the false positive rate (1-Specificity). The true positive rate is the probability that you diagnose a disease, if the patient has the disease, and the false positive rate is the probability that you diagnose that the disease is present if they do not have the disease. For medical purposes, the true positive rate being higher tends to be more important than false positive rate being lower, falsely diagnosing someone with cancer is preferable to failing to diagnose someone who actually has cancer. For example, according to this curve for Edema if we allow a false positive rate of 20% (1-Specificity) then we would expect around an 80% chance of true positive rate. The ROC curve expresses this trade off, with curves that form more of a bend into the top left corner have better performance.

The AUC, shown in Table 2, stands for area under the curve, which represents a type of sum over the ROC. The idea is, the AUC is at minimum 0.5, and at maximum 1.0, with an AUC of 1.0 being perfect diagnosis.

5 Conclusion

References

[R. Selvaraju and Batra, 2019] R. Selvaraju, M. Cogswell, A. D. R. V. D. P. and Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, pages 336–359.

[Wang et al., 2017] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common

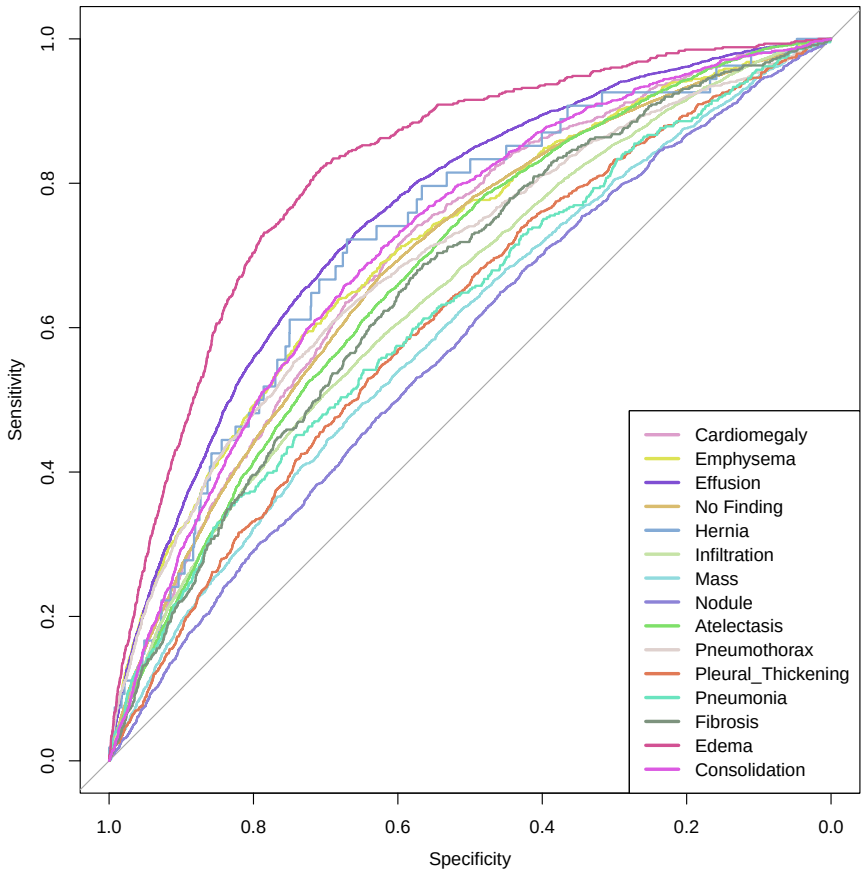

Figure 1: ROC Curves for each class.

Diagnosis	AUC
Cardiomegaly	.701
Emphysema	.709
Effusion	.754
Hernia	.724
Infiltration	.648
Mass	.598
Nodule	.574
Atelectasis	.680
Pneumothorax	.692
PT	.620
Pneumonia	.626
Fibrosis	.661
Edema	.820
Consolidation	.717

Table 1: The AUC for each class.

thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471.

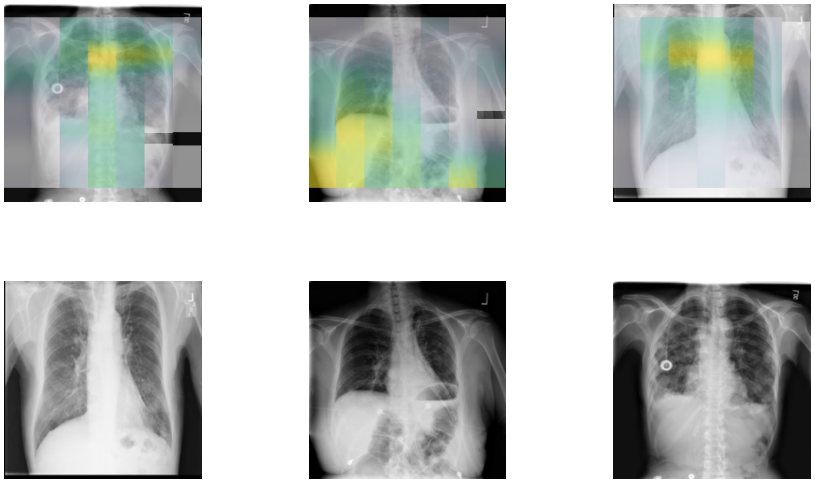

Figure 2: Three different X-rays with heatmaps overlapped showing pixels of interest. The class for each is Pleural Thickening (left), Pneumothorax (center), and Mass (right).