Chexnet2: End to End Improvements for Chest Disease Classification

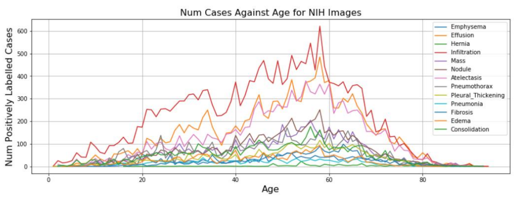
Liam Neath, Alexander Kucy {liamn, akucy}@stanford.edu
Department of Electrical Engineering, Stanford University

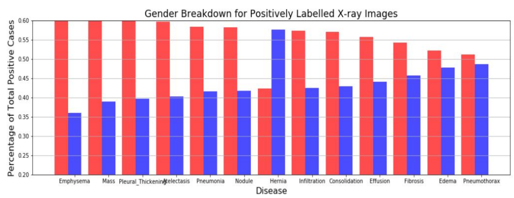
Introduction

Accurate and timely diagnosis of medical images continues to receive substantial academic and industry attention. More recently, the National Institute of Health (NIH) released a substantial amount of labelled chest x-rays which in turn led to the development of a classifier named Chexnet. In the spirit of developing a true medical device, we developed and tested a myriad of techniques, including energy localization, stacked binary relevance and chi-squared informed pairing, aimed at improving classification accuracy within the N.I.H dataset and in an out-of-sample Shenzhen X-ray dataset. Although performance on the N.I.H dataset remained constant, out-of-sample abnormality classification improved by 10%.

Data

The N.I.H provides 112120 grayscale images (1024 x 1024), labelled for 14 common chest diseases and associated with patient age, capture position and gender. The Shenzhen chest x-ray dataset contains 662 frontal chest X-rays, labeled for the presence of tuberculosis, patient age and gender. Shenzhen images have significantly higher resolution at 3000 x 3000 pixels and include spatial information about the location of the disease.

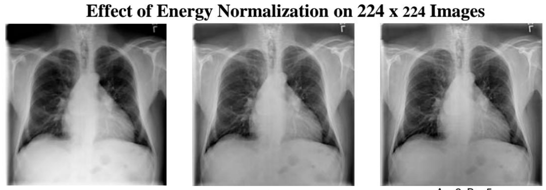




Features

Given the gender and age breakdown of the data, only scaled 224 x 224 images were considered. Scaled images were extremely important in making the training process tractable. A derived normalization energy was also used as a feature given the large variety in image intensity within the datasets.

Effect of Energy Normalization on 224 x 224 Images



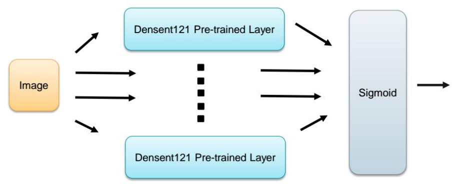
Normal | Λ = 1, B = 5 | Λ = 2, B = 5

Models

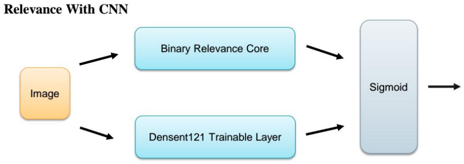
All architectures made use of a base densenet121 neural network architecture with a standard binary cross entropy loss function:

$$L = -\sum_{c=1}^{M} y * \log(p)$$

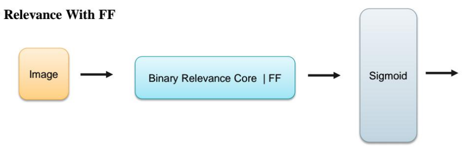
Binary Relevance Model



Binary Relevance With CNN



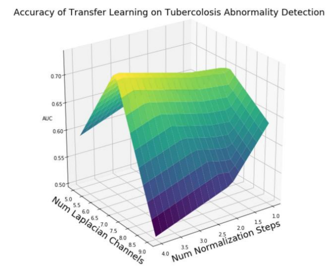
Binary Relevance With FF



Results

Best Test Accuracy of Various Binary Relevance Architectures

	Base Model	Binary Relevance	BRelevance + CNN	BRelevance + FF
14 Class Average	0.839814	0.790908	0.838806	0.685054
Atelectasis	0.817666	0.783850	0.818333	0.781767
Cardiomegaly	0.909232	0.905944	0.899742	0.672409
Effusion	0.883482	0.829249	0.885067	0.837267
Infiltration	0.714012	0.702872	0.750767	0.698776



Discussion

Though deep neural network architecture for medical imaging today is rather standardized, we were able to engineer, through experimentation, several additional features to further improve diagnostic accuracy.

Energy normalization increased accuracy by 10%. As a preprocessing step, not dependent on the size of the data set or computational power, it is low hanging fruit for implementation. Class aware stacking based on chi-squared informed pairing of diagnoses yield 6% greater accuracy.

Our experimenting with binary relevance showed that marginal dependence among images is not captured well, and for classes with a low number of positive examples accuracy greatly suffers. Better to stay with the fully connected network.

Future

We would like to deploy a localization feature using the Grad-Cam method. This would allow us to see how our models weigh regions of the image as important to the classification.

Link to video

Youtube: https://youtu.be/iQqlacwBtUM