Capstone 2: Machine Learning as a Rapid Radiological Screening Tool

……Application of a CNN to chest X-rays for image classification

Milestone Report

I. Background, Problem, and Goal

Many thousands of x-rays are performed daily. Within the medical community the assessment of an x-ray especially a chest x-ray (CXR) , at least in the rapid screening since is a highly common if not ubiquitous task amounts many medical specialties. Official review is done by a Radiologist (MD or DO). To them this task is bordering on trivial in most cases due to the fact that they have evaluated so many. However, to get to that point they have had to undergo many years of school and residency. In addition to this despite being a minor task the shear number of them makes them require time which taxes the system.

Thus, the idea of an automated system that could take on some of the burden. At the least flag potentially, worrisome scans and eventually if it becomes consistent enough relied upon as an autonomous system for medical professional to simply check the results of. The NIH themselves state that they envision a "virtual radiology resident" i.e. a system that can screen x-rays and report to a senior radiologist for confirmation. Another and possibility more pressing use for something like this comes from areas without access to a radiologist (rapid or at all) either a rural area with only a small clinic.

Given that US health spending represents 18% of the GDP and there is a ubiquitous movement to find ways to increase efficiency, decrease costs, and reduce medical errors. There are many potential clients for such a service from HER companies or directly to hospitals. The overall best implementation would be direct integration into an existing EHR system such as EPIC or Athena so that the results could be presented right alongside the actual scan and directly linked to the associated patient medical record for efficient review. Medical professional already has automated warning built into certain EHRs to inform them about a patient being overdue or in a risky range for lab values. The ultimate goal in my mind for a system such as this would be for it to integrate seamlessly into an existing or many existing electronic platforms for ease of use. Rather than simply adding to the complexity and requiring the use of yet one more piece of technology that hinders interaction with patients.

II. The Data

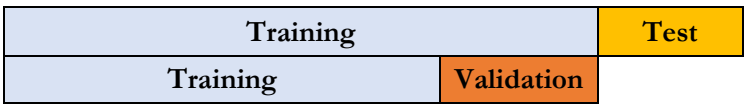
The data for this project originates from the Nation Institutes of Health (NIH). It is a publicly available set of x-ray images and associated labels with some cursory meta-data e.g. age and gender. The data set consists of 112,120 images from over 30,000 patients.

As I will be taking a deep learning approach to this problem and it is a "curated" dataset there is little in the way of classical data wrangling to be done. There are no missing values present within the data and all images have the requisite labels.

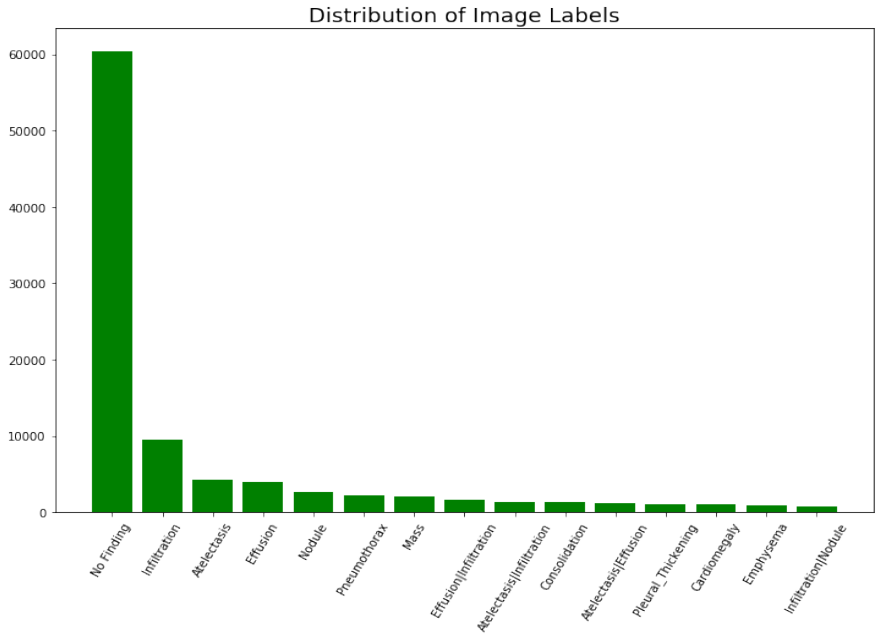
Some issues do exist however. Over half of the images fall under the no finding category leading to some imbalance from the start. Also, as seen below we have moderate imbalance among the individual label categories as well. While this is somewhat representative of reality i.e. certain findings will be more common than others due to higher incidence.

The issues are addressed as follows:

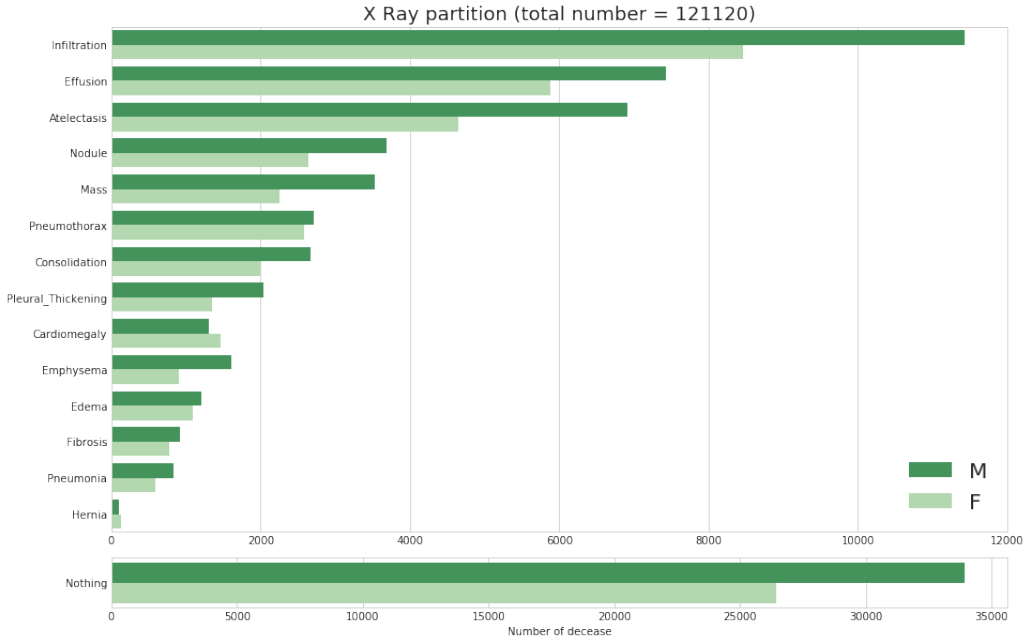
1) Multiple category images will be will be entered more than once under each pathology label for which they are reported (multi label model, see below).
2) Only 9 label categories will be used due to sufficient sample size e.g. hernia has very few samples.
3) The data may have to be balanced if to reduce bias.
4) As the images are currently in one giant pile as it were and through Keras we have the option to 'flow from directory' when pre-processing the images, I will be sorting the images into subfolders based upon the provided labels. These subfolders will then directly provide the relevant label to the model as it runs.
5) The data will be split into 3 groups. First the entire set will be separated into a 80/20 split for training and testing. Then the training data will be split again in an 80/20 manner for training a validation set for use while training the model.



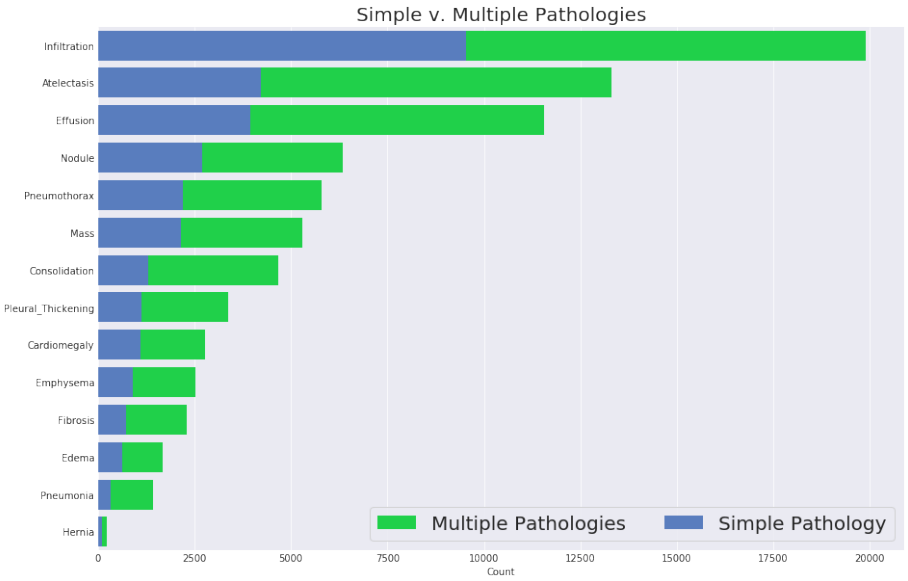
EDA



Here we can see the top 15 labels from the data set. There are further labels that did not make the count into the above table as well. I will be setting a sample cut-off (probably 1000 samples) in order to streamline the model and insure adequate number for both training and testing. Deep learning in general tends to require a great many examples to learn from and if insufficient training samples are provided it may lead to inaccurate results.

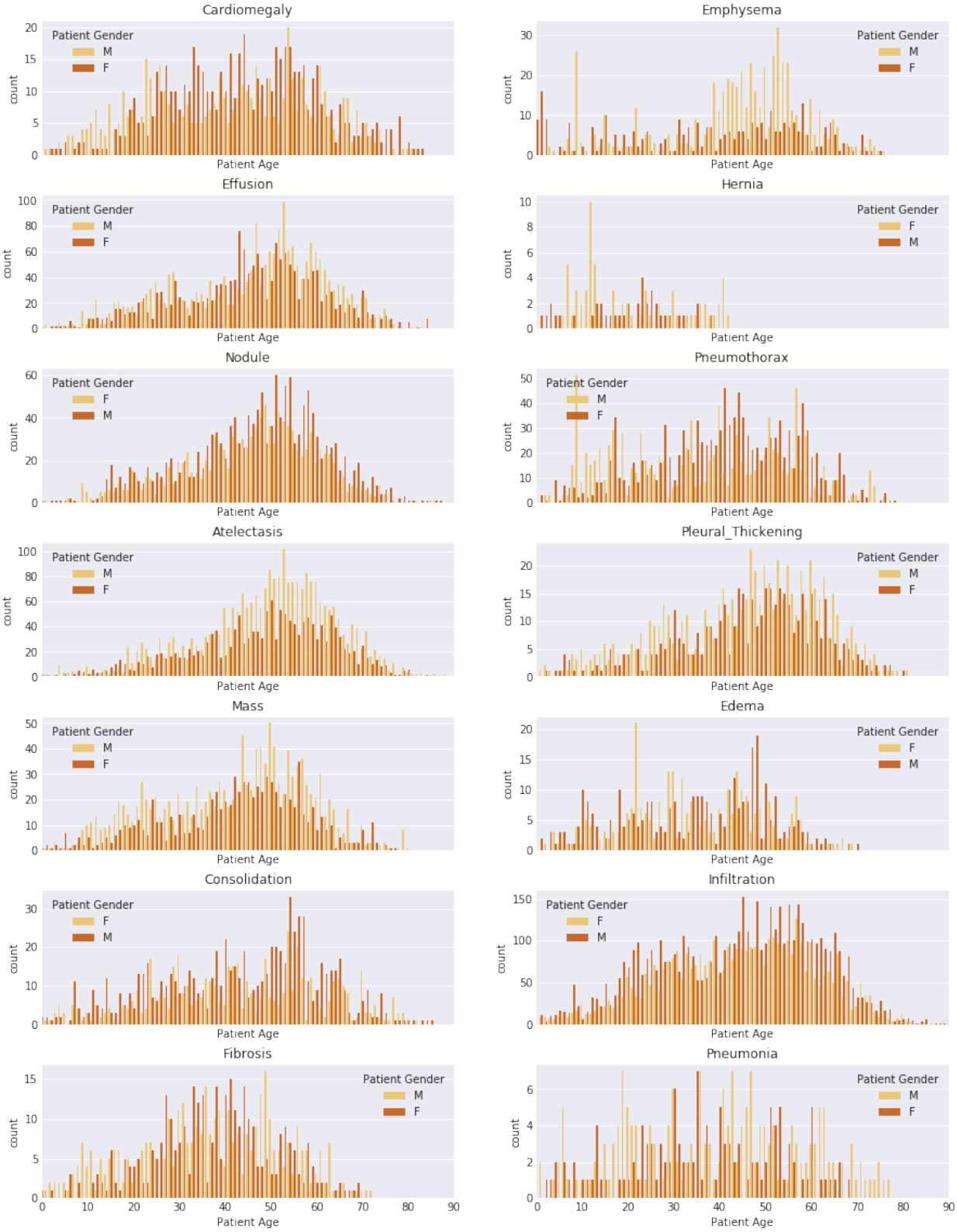


Here we can see the proportion of each label category divided by gender. We do not have an even split but it is roughly balanced for most categories and for these purposes gender should only play a relatively minor role in the image differentiation. Gender, age, and medical history would however play a significant role as far as meta-data that would likely greatly increase the accuracy of more in-depth models.



Here we can see the proportions of samples with multiple vs single label categories. I intend to use a multi label model so this should not be a huge problem. However, it does have the potential to create a biased model. Father, as performed in this paper on this same data set I will only be using 8 pathological labels plus normal/no finding for analysis. This is due to many factors but number of samples for each category and ability for the system to differentiate between some more fine detail diseases make this a more sensible course than trying to build a model for all pathologies in the label list. One of my central concerns for this project is a model than can spot normal (no finding) vs pathology (one of the labels above) very well. The ability to precisely identify the pathology is important but secondary to the creating false negatives (missing pathology).

As can been seen from these charts there are not any major problems with regard to gender or age distribution throughout the dataset in general or in the most of label categories. The is some inconsistency in the Hernia images. However, this is not a concern here as I plan to drop the Hernia category due to insufficient number of samples.


Cardiomegaly
Emphysema
Effusion
Hernia
Nodule
Pneumothorax
Atelectasis
Pleural_Thickening
Mass
Edema
Consolidation
Infiltration
Fibrosis
Pneumonia
Patient Gender
M
F
count
Patient Age

Inherent problems to this kind of data

I. Labeling

Along with the release of this data set a paper was released describing a study conducting an analysis similar to the one being done here. They also describe the creation of this dataset. From this we now that the labels for the images here were obtained using NLP and are estimated to be about 90% accurate. This can be a tricky area even in medicine as many radiologists will disagree on routine diagnosis anyway. Some studies have shown that **human radiologist will outright miss up to 30% of positive findings**. Further studies have shown that there is **disagreement between the interpretation of X-rays in up to 56% of cases.**

Normally in cases such as this an inter-observer score will be used to increase the validity and balance of results. This is beyond the scope of this project. And in the interest of expediency the provided labels must be trusted in this case. An examination of some problems with these labels can be found here (methods for improvement and making a professional grade system will be discussed later).

As a side note the stats above are not meant as a direct critique on radiologists. They are merely to put the current undertaking and proposed tool into context. Also, to point out that, practically speaking, any such tool that is produced would not have to be perfect in order to be useful and improve the quality of care.

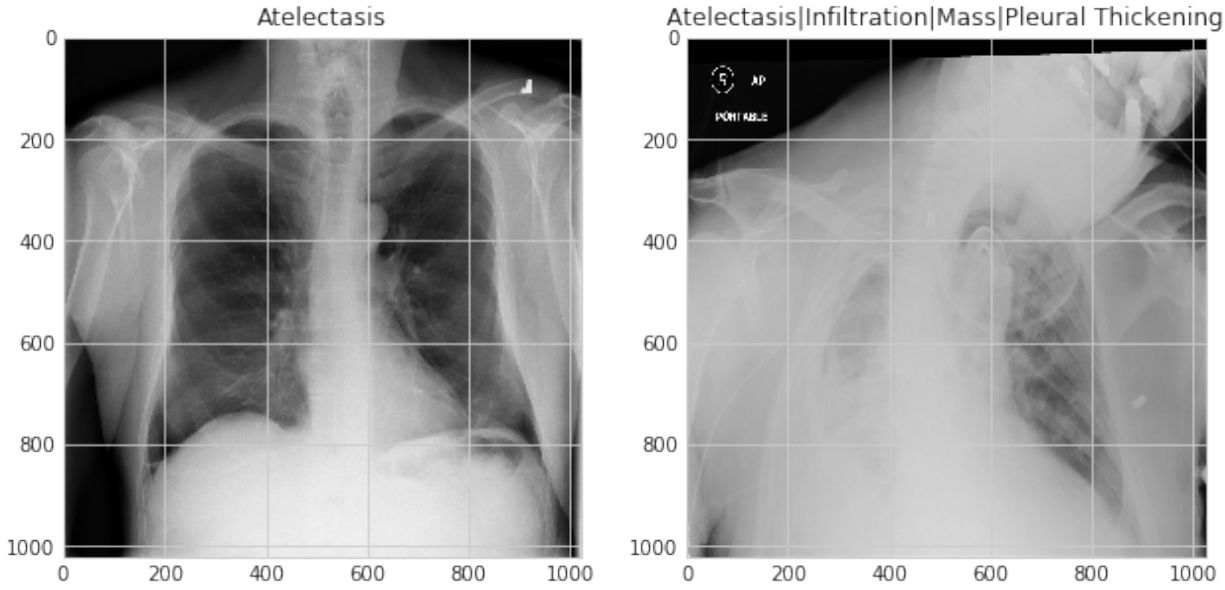
If you are not a medical professional I do not expect you to be able to look at the images below and see much in the way of meaningful data. It takes lots of anatomy, physiology, pathology and plain old experience to be able to reliably interpret these scans, but I would like to point out a few things that might hinder a CNN given the way that it will be processing the features that it reads in the images.

II. Medical Equipment

Many of the images have medical equipment such as various tubes and wires which are a standard part of medical care. Given the pattern recognition manner in which these models work I have some worry about it coming to associate such devices with a disease process rather than the anatomical changes seen in the image. For example, as discussed here the model may come to

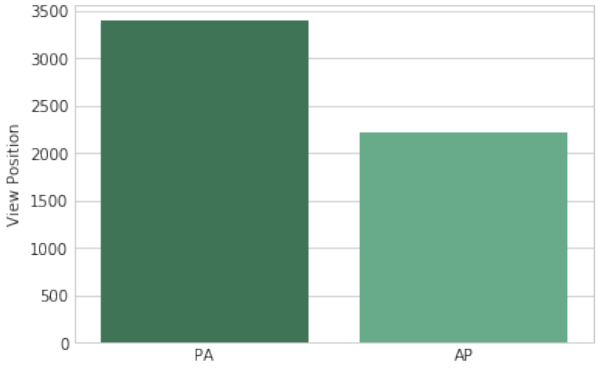
recognize a chest tube as associated with a pneumothorax. Such an association may not be entirely bad but one does not necessarily imply the other and we would need the model to recognize a pneumothorax without the presence of chest tube.

III. Simple vs Multiple Pathologies

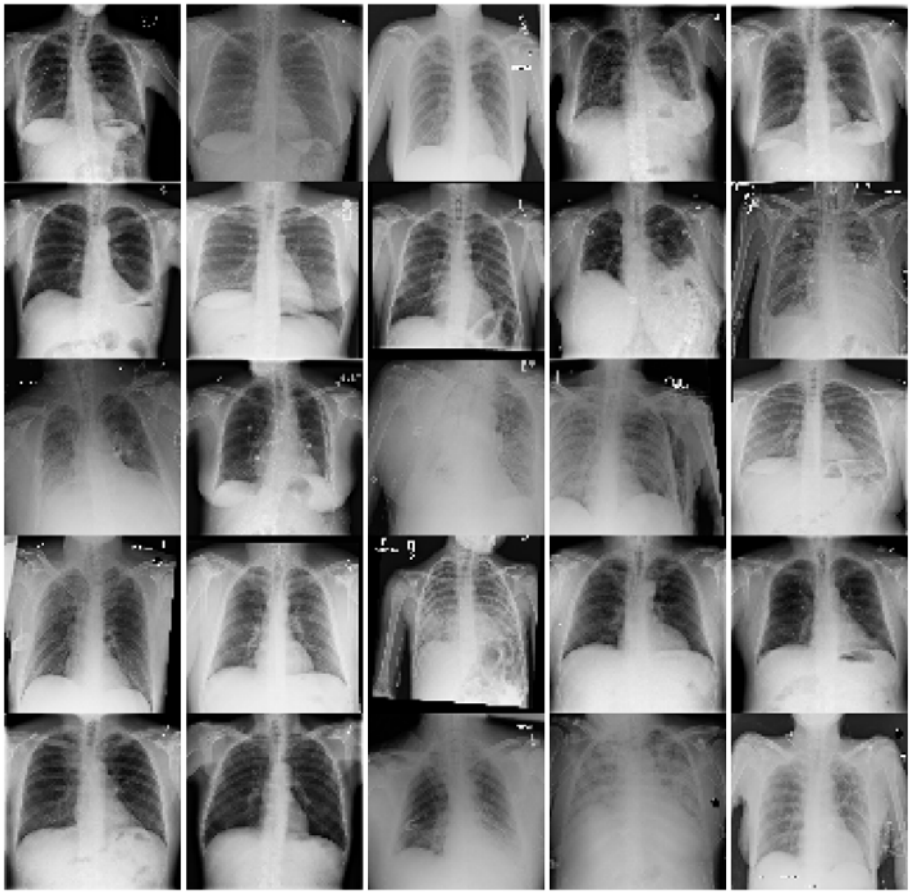


These two images demonstrate several potential problems to deal with the image on the left has a disease process called atelectasis, or partial collapse of a lung segment. The image on the right has several pathological labels. However, it also has an impaired viewing angle compared to the left due to patient position when the image was taken. The x-ray on the left represents just about a perfect view with clear representation of all areas that need to be evaluated by a human when examining an x-ray. The image on the right represents a not too uncommon lack of those things.

IV. View Angle



The anteroposterior (from the front) and the posteroanterior (from the back) X-ray views are both common and differing aspects are sometimes more easily seen from one view over the other. The PA is the primary view used with the AP acting as more of a supplemental view in most cases. The AP view has a tendency to alter the proportions and/or obscure certain structures and thus requires a greater degree of technical expertise to accurately interpret. Any successful model would need to seamlessly evaluate both views. However, I would anticipate that it could be possible for error to be induced if the model is evaluating the same disease process from two different views.

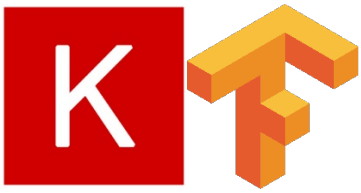


In the past google and others have demonstrated that simply using 'messy' data a brute forcing lots of it harnessing raw computational power can work as well if not better than heavily curated data. In theory this might could work for a model such as this allowing us to lower the threshold a bit on what it considered unusable data. I do not know if this is such a use case and if it is I would theorize you would need a couple million images rather than 100k. Not to mention a great deal more training time than I will be able to apply to the model discussed here.

III. Methodology and model

a. **Use of CNN** (why and how)

CNNs are currently the mainstream model in use for image recognition.

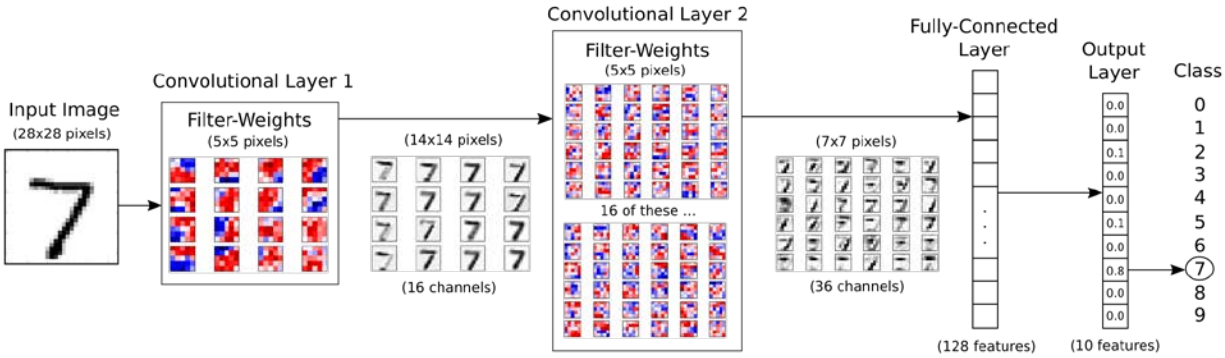


For my purposed I will be using Keras with a Tensor Flow backend. The current iteration of the model uses the sequential model, with 'ReLu' activations and 'MaxPooling' layers. Dropout and SoftMax layers are also employed. The exact structure, influence and interaction of individual layers, and training dynamics will be discussed in detail in the final write-up for this project. At present using an "out of the box" model I obtained a 55% accuracy when classifying the images across 9 categories.

Running this data will not be like playing around with the MNIST data set. The MNIST data consist of 60,000 training and 10,000 test images that are 28x28 pixels. The size of all combined files comes to less than 20Mb. This data set can easily be run on just a modern CPU. The data being used here consist of 112,120 images that are 1024x1024 pixels (already scaled down). The complete data set comes out to around 42Gb. Given this, I will be using GPU processing in a cloud environment. **Note:** for model building and initial testing a smaller randomized and categorically proportional data set consisting of about $1/5^{th}$ of the larger set was used.

The current standard for cloud computing is Amazon Web Services (AWS) and while this is the most feature rich and robust option generally speaking I will not be using it for this project. There many options but in the interest of simplicity I will be using Floyd Hub (alternate option is Paper Space). This is a service built on top of AWS which has pre-set instance types with CPU and GPU options along with storage. They also provide pre-set environments e.g. TensorFlow with Keras, Caffe, and Pytorch with many of the required and useful dependencies already installed. The use of this service is a convenience not a requirement.

As an initial exploratory step and to obtain a ballpark range of what to expect I created an initial model (along the general structure of what is shown below). This was done only with minor expectations and more as a building block than anything final.



The specific structures of the model, its results, generalizability, and fine tuning will be discussed in detail in the final report for the project.

IV. Next steps

"The idea of knowledge is cumulative —seeing farther by standing on the shoulders of giants"

- Isaac Newton

From this point the plan is to 1) apply a transfer learning model using a pre-trained model and adapting it to the current data. This will likely be done using either the ResNet50 or similar models. These models have already been heavily trained on very large image sets and in theory should greatly increase the robustness of the results here without having to spend exorbitant numbers of hours directly training this model from scratch. 2) Once the new model is working and tuned I will then apply it to the full data set. As mentioned above this is not being done directly during the exploration phase in order to conserve computing resources.

V. Resources

https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

https://medium.com/intuitionmachine/the-brute-force-method-of-deep-learning-innovation-58b497323ae5

https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/

http://www.radisphereradiology.com/wp-content/uploads/Diagnostic-Accuracy-in-Radiology