Machine Learning Engineer NanoDegree

Capstone Proposal

Name: Rishabh Sobti
Date: 4th February, 2018

Domain Background

This project aims at inclusion of machine learning technology into the field of medical diagnosis. One very important part of lung and thorax assessment is chest X-Rays. The proper study of chest X-Rays can help doctors diagnose several diseases like effusion, infiltration, masses or nodules in the lungs, etc. As beneficial as chest X-rays are, they are also challenging to read and infer properly. A lot of work is already going on, in the field of computer-aided detection and diagnosis, but achievement of a satisfactory level of accuracy in such a delicate field is a dream yet to be achieved. A promising example in the field is using CNNs for detection of Tuberculosis (source: CNN for TB).

With the increasing amount of chest X-ray data openly available at our disposal, deep learning techniques can be very useful to study trends and aid doctors in reading chest X-rays. I come from a country where 5.2 million medical errors are happening annually, most of which are due to wrong diagnosis, so, I feel motivated to work in the given field. It should be noted, that ideology is not to eliminate the involvement of doctors, rather to help them make the right diagnosis.

Problem Statement

The goal is to create a system which can classify input X-ray images into the 15 classes (14 diseases and one class for “no findings”), the details of which are discussed in the next section of this proposal. The tasks that are needed to be performed are:

1. Downloading and preprocessing the needed dataset.
2. Designing a convolution neural network, which can train on the given data.
3. Making the model optimal, by adjusting parameters, that leads to a higher accuracy.

In logical terms, it is an image classification problem, and since the dataset, which will be used is labeled, it is both quantifiable and measurable, with help of metrics like F-score.

Dataset

The dataset hosted by the National Institutes of Health – Clinical Center (official website - https://clinicalcenter.nih.gov), is the most suited in order to achieve the desired goal. It is a collection of 112120 chest X-ray scans, from 30805 unique patients, each of 1024X1024 resolution, which are labeled into 15 classes. The dataset is openly available for download and use at "https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345".

But, due to lack of computing capacity at my disposal and the time limit on this project, I will be using a drastically and randomly reduced version of this complete dataset, which is contributed at and reviewed by kaggle.com. The link for the sampled dataset that I would be using is https://www.kaggle.com/nih-chest-xrays/sample . The dataset can be obtained using this link, after going through a free signup. To make the downloading process even easier, I've uploaded the same dataset at this google drive link – https://drive.google.com/open?id=1IuCB5Etj-yOVrVtvM6Y8rybJKn4LCX3d . The labels on these images are provided in a separate .csv file that accompanies the dataset. The 15 disease labels used in this dataset, as mined using Natural Language Processing on the original medical reports are:

1. Hernia - 13 images
2. Pneumonia - 62 images
3. Fibrosis - 84 images
4. Edema - 118 images
5. Emphysema - 127 images
6. Cardiomegaly - 141 images
7. Pleural_Thickening - 176 images
8. Consolidation - 226 images
9. Pneumothorax - 271 images
10. Mass - 284 images
11. Nodule - 313 images
12. Atelectasis - 508 images
13. Effusion - 644 images
14. Infiltration - 967 images
15. No Finding - 3044 images

The size of this reduced dataset is ~2GB, in comparison to 45.6GB dataset originally there. It contains a total of 5606 images of the resolution 1024X1024.

Solution Statement

The solution to the problem addressed above is:

1. Designing a convolution neural network, which consists of several convolution layers and pooling layers, followed by fully connected layers.
2. Adjusting the parameters associated with these layers, in order to maximize the prediction F-score of the given CNN.

The given dataset will be divided into testing, training and validation subsets, in order to train and validate the neural network. The solution or the goal will be achieved when the metrics seem promising enough, and then this project can be extended to be used on the complete dataset.

Benchmark Model

I will be building and comparing my model in the following two phases:

1. Construction phase: While constructing and deciding the number and type of layers of the convolution neural network, I will aim to get F-score anywhere above the F-score value of a vanilla neural network on the same dataset. The detailed analysis of the simple vanilla NN model, with only one hidden layer, will be incorporated within the project by me, and will be used as the basis in construction phase.
2. Optimization phase: While optimizing the parameters and connections of the CNN, I will attempt to reach an F-score of more than 0.5. The definition of F-score is given in the next section. Since, I am using only ~5000 images for classification into 15 classes, with half of the classes having only sparse values, 0.5 is more than reasonable F-score to expect. This is what I have judged by observing the outcomes of several CNN models online.

Evaluation Metrics

Since, we need to make a model which correctly predicts the disease associated with the chest X-ray, and the dataset we have is imbalanced, we need to consider F-score for the model, which gives a fair idea about the model if the dataset is sparse. The F-score can be defined as the harmonic mean of two other metrics, which are recall and precision.

Mathematically, for each class, we can calculate recall and precision as:

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Where TP: True positive or the samples of this class which are predicted correctly

FN: False negative or the samples which belong to this particular class, but predicted incorrectly

FP: False Positive or the samples which do not belong to this class, but are predicted to be in this class

And, for each class, F-1 score or simply, F-score is:

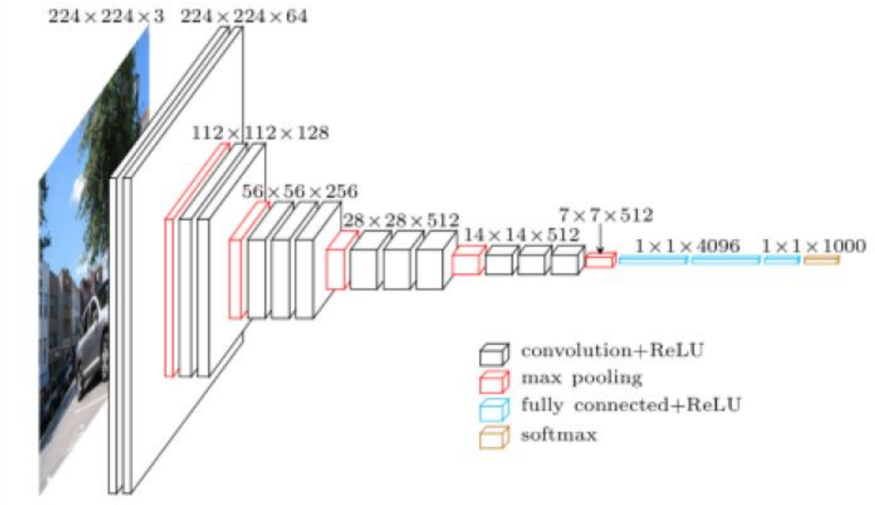
$$F - score = 2 * \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}}$$

The F-score of all the classes are then averaged in a weighted fashion, to get F-score of the model.

Project Design

The outline of the design I intend to use for this project is as follows:

1. First, I'll be extracting the data, and visualizing it, to get theoretical insights into the statistics of the given dataset.
2. Then, I'll normalize the data, one-hot encode to the labels, and partition the training, testing and validation subsets.
3. After data preprocessing, I'll implement a simple 1-layer vanilla neural network to set a basic benchmark for my model.
4. At first, I'll be feeding the images in their original size for training the CNN, because resizing them to lower resolutions may lead to loss of sensitive information in the chest X-Ray, but, if the computation overheads are too large, I'll consider resizing them to lower resolutions.
5. Then, I plan to implement my initial structure in the following way (the dimensions of my image and model will be different from what is shown here):



(Source of the image: https://flyyufelix.github.io/2016/10/08/fine-tuning-in-keras-part2.html)

After these steps, I plan to adjust my model parameters and number of layers based on the F-score I obtain from my initial setup.