Chest X-ray Abnormality Detection
Capstone Project

Problem Statement

According to the Center for Disease Control (CDC), on average, 48 out of 100 people who will visit a hospital will receive an x-ray. Some of these will be in relation to a pre-identified condition, and others will be a diagnostic measure to confirm the presence or eradication of a condition. The latter of the two is an increasingly important measure to diagnose a possible ailment that could be life-threatening and it helps reduce total healthcare costs for individuals and the nation at large.

However, chest x-rays are infamously inconclusive and as we age, anatomical changes can cause even further ambiguity among results, as noted here. In cases like that, consensus matters, which is why improving the accuracy of identifying any given condition is important. Early and accurate detection reduces time spent in a hospital as well as visits, reduces the need for an invasive procedure, and it would also provide invaluable information regarding disease progression or regression.

From an x-ray, if a radiologist were able to identify an observed abnormality, and work in conjunction with a computer aided detection (CAD) algorithm that could tell with almost 100 percent certainty that the condition observed by the radiologist was indeed present or would need another opinion, this would greatly reduce hospital costs for patients and staffing issues for the hospital's managers. It would also be of great impact to developing countries where they may have one machine for every million people, because that one radiologist could tell with validated certainty what five radiologists would be needed to identify.

Data Wrangling and Cleaning

The data was provided through Kaggle's competition: **VinBigData Chest Abnormalities Detection**, and was curated by the VInBigData Institute. The 18000 unique chest xray images were provided to the VinBigData institute by 2 hospitals in Vietnam: the 108 Hospital and the Hanoi Medical University Hospital. In total, there were 4 datasets, and they are: 1) submission

csv file, which was meant to provide an example of a properly formatted inference result. 2) A csv file containing the training data, which has a total of 67914 observations and 8 columns. Even though there are only 15000 unique images, the totality of this csv file comes from multiple notations and abnormalities detected on some of the images, and because for each image, three radiologists made individual annotations as to the presence of an abnormality, or not. In other words, an 'image_id' will be observed and recorded a minimum of three times, and if it possesses multiple conditions, it may be annotated a corresponding amount of times. 3) A file path containing all of the training images stored with a .dicom extension, totaling to 15000. 4) A file path containing all of the testing images stored with a .dicom extension, totalling to 3000.

The datasets were clean for the most part, but within the csv file containing the training features, all rows that possessed a 'class_id' of 'No finding', had NaN values for their bounding box coordinates. And since roughly 47% of the observations were of 'No finding', this required attention. By simply filling the missing values with 0, the model would function properly. This however, was a symptom of a problem encountered later through exploratory data analysis, that is, this data was massively imbalanced.

Exploratory Data Analysis

A bit about DICOM

DICOM stands for Digital Imaging and Communications in Medicine. Within the healthcare industry, and hospital setting in particular, the recording, transferability and interoperability of patient profiles has been a difficult task to standardize. Wikipedia has a great explanation here, but in the early 1980s, DICOM profile formatting was created to tackle this issue. These files consist of metadata ranging from a unique patient-identifier key, age, sex, information as to what image is being stored, i.e."Digital x-ray image", rows and columns of the pixel data, its bits stored quantity, and also its transfer syntax, which represents the type of encoding used to store this image for recreation and manipulation across platforms. There are some great Python libraries to work with DICOM images, like pydicom, which facilitates intuitive functions to access a DICOM formatted file's metadata. In the setting I worked under, extracting the pixel data to recreate and manipulate the size of the image was the only necessary step to take.

There are 15 classes in total, composed of 14 diagnosed conditions and one class of "No finding". The following provides a brief overview of the 14 (0 indexed) diagnosable classes.

0 - Aortic Enlargement:

- Formally known at an aortic aneurysm, occurs when there is an abnormal bulge in the aortic wall, causing it to enlarge or dissect.

1 - Atelectasis

- A complete or partial collapse of the entire lung or an area of the lung. It occurs when the alveoli within the lung become deflated or filled with alveolar fluid.

2 - Calcification

- The accumulation of calcium salts in a body tissue.

3 - Cardiomegaly

- An enlarged heart.

4 - Consolidation

- Refers to the presence and solidification of exudate in the airways and alveoli, usually as a result of infection. When present in the lungs, pulmonary consolidation is referred to as pneumonia.

5 - ILD

- Interstitial lung disease (ILD) is a group of many lung conditions that affect the interstitium, which is a network of tissue supporting the lungs' alveoli.

6 - Infiltration

- Pulmonary infiltrates are substances denser than air, such as pus, blood or protein which lingers within the parenchyma of the lungs. Responsible for pneumonia and tuberculosis.

7 - Lung Opacity

- Observed through a radiograph or CT scan, refers to a decrease in the ratio of gas to soft tissue, resulting in a darkening of an area of the image. Represents the presence of a liquid or disease in the lungs.

8 - Nodule/Mass

- Small masses of tissue in the lung.

9 - Other lesion

- An area underneath the skin that exhibits abnormal growth, not identified as a nodule/mass.

10 - Pleural effusion

- The build-up of excess fluid between the layers of the pleura outside the lungs. Pleura exist between the lungs and the chest cavity, helping to facilitate breathing.

11 - Pleural thickening

- Develops when scar tissue thickens the pleura.

12 - Pneumothorax

- Collapsed lung, occurs when air leaks into the space between the lung and chest wall.

13 - Pulmonary fibrosis

- A lung disease that occurs when lung tissue becomes damaged and scarred. Progresses by preventing the lungs from properly functioning.

Some of the labels exist within a similar capacity to one another. In other words, it may be likely for a radiologist to observe a nodule/mass and/or a lesion. It may also be likely for a radiologist to observe and label the condition as pneumothorax, when in fact it was atelectasis. It may also be likely that a radiologist observed and labeled a condition as a lung opacity, when it could have been a consolidation or infiltrate that was present. The point being, many of the classes represent a general classification for what may be going on, but may not be conclusive. Without an invasive procedure, it may be impossible to conclusively determine what condition is present, which is why this dataset and competition is so important. If there were a way to detect these objects with near perfect accuracy, it would save millions in medical procedural costs, and prevent unnecessary invasive procedures.

Preprocessing

The pixel arrays extracted from each DICOM file were shaped differently, so rescaling was a necessary step. Then once rescaled, each image was resized to dimensions of (256, 256). ImageNet-scale networks can accommodate images of this size, and although this dataset is not as large as an ImageNet competition, it's important to become familiar with large-scale frameworks. Labels and corresponding bounding boxes were extracted from the training csv file, and formatted into stacked numpy arrays for input into the keras CNN using tensorflow-backend. FInally, the labels were encoded usinge sklearn's MultiLabelBinarizer to create same-length arrays of label-input data to facilitate the multilabel outcome.

Model Selection

For object detection tasks, CNN's (Convolutional Neural Networks) are the best performing architectures, with iterations like Faster Regional Convolutional Neural Networks (Faster R-CNN) and Mask Regional Convolutional Neural Networks (Mask R-CNN) being the state-of-the-art algorithms for handling such tasks. I chose to implement a basic CNN for this task to focus on the multilabel classification and single bounding box predictive power of CNNs. The caveat to this implementation of a CNN was that there are two output heads stemming from the convolutional layers. In total, there are six convolutional layers, and each output head contains 3 dense layers. Normalizing the data after sets of convolutional layers improved the performance of the model, and after each normalization layer, a max-pooling layer was also implemented to limit parameters.

The best label classification score I have received thus far through validation (using .evaluate() method from keras) was 92%, with a bounding-box accuracy of 72%, and the best mAP (Mean Average Precision) score I have received is 0.22 on the complete set of classes.

Such low label and bounding box accuracy, coupled with a poor mAP score leads me to believe the imbalance in data is causing the model to be overfit to one class ('No finding') and underfit on the rest. To account for this, implementing a L2 regularization penalty may help or replace normal Dropout layers with SpatialDropout layers.

Takeaways

A basic CNN has its limitations. When it comes to classification tasks, specifically for images, it is considered to be near state-of-the-art, but when it comes to object detection, without conducting selective search for region-of-interest (ROI) pooling, it is unable to detect multiple objects within a given image. There are frameworks that exist that utilize CNNs for object detection, like R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN, which all use some variation of selective search and ROI pooling before or while inference is taking place, and these deep learning architectures continue to push the boundary of near real-time inference. In the developed world, fluoroscopy is becoming the standard diagnostic routine, so

realtime analysis is even more valuable, but in the developing world, image classification and object detection of a still image would still be life-changing technology.

This model is the first iteration of what will become the backbone for a progressively deeper model, capable of conducting a full multilabel, multi object detection task. For a client to utilize this model and analysis, I recommend 1) utilizing it as an auxiliary test/confirmation of an observed condition. With further iterations and expansion of trainable data, this could become a backbone for a company's analysis software. 2) Quick analysis could extend the diagnosis capabilities of a single radiologist or a single physician if they had this assistance. In areas around the world where access to healthcare professionals is limited, and access to diagnostic equipment is even more limited, having a structure like this algorithm to assist in identifying what may be present in a CT scan will lighten the burden placed on an already stressed healthcare infrastructure. 3) Implement this algorithm in regions around the world. The dataset that trained this model was from one country, so training 'regional' models and then drawing from the consensus of inferences could expedite diagnoses and provide a much more balanced prediction. For example, cardiovascular disease is a leading cause of death in many countries, but not others. So having a model that is near perfect at identifying the nuances with a chest x-ray to identify a condition, could be incredibly helpful for a country that is starting to experience a rise in such conditions. This would help keep that locality ahead or at least in-step with a rise of a condition or disease, because detecting early-onset is a first-line of defense when it comes to lowering healthcare costs for any economy.