Machine Learning Engineer Nanodegree Capstone Proposal

Pablo R. Lopez
May 20th, 2021

Medical Diagnosis: Pneumonia Detection from X-Ray images using Computer Vision.

Domain Background

Pneumonia is an inflammatory condition of the lung primarily affecting the small air sacs known as alveoli. Symptoms typically include some combination of productive or dry cough, chest pain, fever and difficulty breathing. The severity of the condition is variable. Pneumonia is a form of acute respiratory infection that affects the lungs. It is caused by viruses, bacteria and fungi and it can be treated with antibiotics.

Therefore, in order to alleviate this problem doctors are relying more in clinical decision support algorithms as a diagnostic tool. Image classification is an area of Deep Learning where the field of Health Care applies Deep Learning architecture models in disease detection into practical applications. Convolutional Neural Networks along side Computer Vision could aid in diagnosing Pneumonia using chest X-ray images.

Problem Statement

The project objective is to use X-ray images to create an algorithm to identify patients with pneumonia vs normal. Therefore, this is a Computer Vision and classification task.

The dataset seems to be about 1.2 Gb in size and this raises the question if is big enough for our algorithm to perform well. Also, the dataset seems to be unbalanced and for the most part it is always best to have an equal number of images for each predicted class when doing image classification. In addition, the images seems to be in a grayscale format. This raises the question if it is necessary to convert grayscale images to RGB in order for the better image object detection.

Lastly, the expectation is to facilitate medical doctors or technicians to upload patient images into a web-app and thus be able to decide if the patient has pneumonia vs normal.

Datasets and Inputs

The original dataset was provided by Daniel Kermany, Kang Zhang, and Michael Goldbaum. This is a part of a larger dataset named **Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images**.

They collected and labeled a total of 5,232 chest X-ray images from children, including 3,883 characterized as depicting pneumonia (2,538 bacterial and 1,345 viral) and 1,349 normal, from a total of 5,856 patients.

This dataset is organized into 2 folders (train, test) and contains subfolders for each image category (Pneumonia/Normal).

Input Data fields

Here is an example of the dataset directory:

```
\chest_xray
    \test
        \NORMAL
        \PNEUMONIA
    \train
        \NORMAL
        \PNEUMONIA
```

Solution Statement

The immediate solution is to build a Deep Learning models such as Convolutional Neural Network (CNN) to find patterns and detect objects to predict which images has pneumonia vs normal. After finalizing the model a front-end serving web-app will be created to load images and determine a diagnosis.

Benchmark Model

For the benchmark model a CNN VGG-16 will be use. The VGG stands for Visual Geometry Group and it is composed of 13 convolutional and 3 fully-connected layers and as well as the ReLU activation function. This network stacks more layers onto AlexNet and use smaller size filters (2x2 and 3x3). It consist of 138M parameters and takes up about 500MB of storage space.

The goal is to obtain an accuracy higher than 90%. As well as, sensitivity and specificity higher than 80%.

Evaluation Metrics

The performance of the algorithm will be evaluated using a confusion matrix table. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class.

	Positive	**Negative**
Positive	TP	FP
Negative	FN	TN

As well as will be using Sensitivity and Specificity.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Sensitivity (true positive) tells us about all the positive cases in the dataset, how many of them are successfully identified. Specificity (true negative) measures all the negative cases in the dataset and how many of them are successful identified. The true positive rate and the true negative rate can be visualize in the ROC curve which evaluate the model's performance of a classification model at all classification thresholds.

Project Design

For the project design we will be using some aspect of the Machine Learning Life Cycle.

- Load data
- Split data into train/test
- Exploratory Data Analysis
- Image pre-processing
- Modeling (CNN) and Fine-tuning
- Model Evaluation and testing
- Deployment (Serving with a front-end)

A quick EDA will be conducted to get a quick look of the data structure (2D vs 3D images), random spot checks in a pixel level by looking at intensity profiles of individual images. We will also check for data imbalance. In the image preprocessing phase will be removing potential noise from images (e.g. background extraction), perhaps perform some normalization, conduct image augmentation and resize images for the CNN architecture.

After, preprocessing the images will run the CNN algorithm and check how it performs. Then, we can fine-tune it by reusing, freezing and adding layers to our pre-trained weights of the first few layers. Finally, after we have our top performance algorithm a front-end will be create it to start making diagnostics on the images.

References

- VGG16 – Convolutional Network for Classification and Detection
- Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images - Mendelay Data
- World Health Organization - Pneumonia