Machine Learning Engineer Nanodegree

Capstone Proposal

François Masson
September 16th, 2018

Proposal

Pneumonia detection based on chest X-Ray Images

Domain Background

The proposed project is based on the attempt of a lung detection and recognition model infected with pneumonia. The idea came by browsing the Kaggle website (link). The essence of artificial intelligence is the amount of data. Element very difficult sometimes to get. While searching on Google, I realized that many other people and scientists had already investigated this problem. Here are some references found on this subject:

https://www.ncbi.nlm.nih.gov/pubmed/25214377

https://stanfordmlgroup.github.io/projects/chexnet/

http://cs229.stanford.edu/proj2017/final-reports/5231221.pdf

It is obvious that recognition of infected lungs is essential as pneumonia is the common type of infection found in the world. The infection spreads in the lungs area of a human body. But it is not always obvious to detect based on X-Ray images. This project is very inspiring for me because artificial intelligence has the ability to process and recognize thousands of images, patterns, structures that a human is not able to do. Medicine is therefore the perfect field of application for this. What fascinated me the most during the hours to study during this program is how you can build CNNs for dog breed recognition and skin cancer detection. I wanted in my way to take this concept and apply them to a different field.

Problem Statement

The problem itself is quite simple and even binary. It comes down to being able to determine if the lungs are infected with pneumonia or not. The problem is therefore quantifiable and an example of output would be a yes or no vector. However, it is very easy to imagine the weaknesses of this system: the system will detect only lungs with

pneumonia but ignore other possible diseases leaving the possibility of having a sick patient but not detected. However, it would be interesting to imagine building a super-model to integrate a maximum of possible diseases to provide a complete scan of the lungs of a patient.

Datasets and Inputs

As mentioned before, the inspiration for this project came by browsing the Kaggle website to get me a database and images. This one is available via the following link:

https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

There are 5863 images in two categories: normal (1583) and pneumonia (4280). There are in a gray scale format but with various size. So, a standardization will be necessary. It appears that the classes are not balanced. This will influence the selection of the evaluation metric. Moreover, it is a relatively correct number to allow to separate these images in training set, test set and validation set. These images were made public thanks to Paul Mooney.

Solution Statement

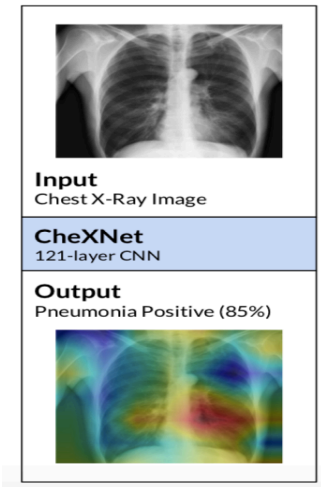
One of the possible solutions for this problem is to use deep learning and the use of a Convolutional Neural Networks architecture. Firstly, by separating the different images into three sets of data, I hope to be able to build an architecture that allows for a correct classification. The output would be to have a percentage of certainty on the highest possible image facing the following question: Are the lungs infected? Yes or no and with what certainty. In a second, the program should be able to bring out the places that seem infected. In addition to being a decision aid, it would be really amazing to be able to locate the infected places or motives leading to a pneumonia decision.

Benchmark Model

By definition the benchmark act as a threshold that can determine if the project has succeeded or not. Therefore, I will train and test a vanilla CNN model. This model is the same as proposed in project associated with chapter 5 of the program:

The number of output layers will be changed and reduced to two. In the same way, the metrics measure will be modified to obtain the F1 score and not the accuracy measure. So, every new model train and tests in the capstone project will be compared to the results obtained with that vanilla CNN model.

For information, the model provided in the article: « CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning » (https://arxiv.org/pdf/1711.05225.pdf) is a 121-layer convolutional neural network trained on ChestX-ray14, currently the largest publicly available chest X- ray dataset, containing over 100,000 frontal-view X-ray images with 14 diseases. As provided below, they get an accuracy 76% for detecting pneumonia in infected lungs. This may get a hint on how it is possible to performed…



CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pathology	Wang et al. (2017)	Yao et al. (2017)	CheXNet (ours)
Atelectasis	0.716	0.772	**0.8094**
Cardiomegaly	0.807	0.904	**0.9248**
Effusion	0.784	0.859	**0.8638**
Infiltration	0.609	0.695	**0.7345**
Mass	0.706	0.792	**0.8676**
Nodule	0.671	0.717	**0.7802**
Pneumonia	0.633	0.713	**0.7680**
Pneumothorax	0.806	0.841	**0.8887**
Consolidation	0.708	0.788	**0.7901**
Edema	0.835	0.882	**0.8878**
Emphysema	0.815	0.829	**0.9371**
Fibrosis	0.769	0.767	**0.8047**
Pleural Thickening	0.708	0.765	**0.8062**
Hernia	0.767	0.914	**0.9164**

Evaluation Metrics

As a reminder, the proposed project is based on the two-stage classification of lung analysis: Normal or Pneumonia. In a first attempt, I considered the use of the accuracy measure. But the dataset classes are not well balanced: normal (27%) and pneumonia (73%). Therefore, the use of F1 score as a measure of metric evaluation is more accurate than the accuracy measure. It will avoid the accuracy paradox. F1 score is defined as:

$$F1\ score = \frac{2 * true\ positive}{2 * true\ positives + false\ negative + false\ positive}$$

Project Design

For the resolution of this problem, the approach will be similar to the one used when carrying out the project associated with chapter 5 of the program.

- Step 0: Import Datasets

- Step 1: Separate Data in train, test and validation data

- Step 2: Create a CNN to Classify normal and pneumonia detection based on chest X-Ray Images (from Scratch)

The data will be pre-processed. The images will be rescaled by dividing every pixel in every image by 255 for example. A possible architecture used will be similar the image below:

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 223, 223, 16)	208
max_pooling2d_1 (MaxPooling2	(None, 111, 111, 16)	0
conv2d_2 (Conv2D)	(None, 110, 110, 32)	2080
max_pooling2d_2 (MaxPooling2	(None, 55, 55, 32)	0
conv2d_3 (Conv2D)	(None, 54, 54, 64)	8256
max_pooling2d_3 (MaxPooling2	(None, 27, 27, 64)	0
global_average_pooling2d_1 (	(None, 64)	0
dense_1 (Dense)	(None, 133)	8645

Total params: 19,189.0
Trainable params: 19,189.0
Non-trainable params: 0.0

The CNN architecture works well for the image classification task because its convolution layers can detect regional patterns in the images and the pooling layers can reduce dimensionality of an array. The dropout layers could be used also because it prevents overfitting, and the Relu functions were used to deal with the vanishing gradient problem. The softmax function is the best to aquire probability for each output node. In this case, the number of output layer is equal to two. Of course, optimizer, loss, learning rate, momentum or others hyperparameters will be tuned to try to get highest accuracy possible.

- Step 3: Use a CNN to normal and pneumonia detection based on chest X-Ray Images (using Transfer Learning)

- Step 4: Create a CNN to Classify normal and pneumonia detection based on chest X-Ray Images (using Transfer Learning)

- Step 5: Test the Algorithm

The goal of this project is to find a way to correctly and with some accuracy recognize lungs infected by the pneumonia. The best is to be able to detect the pattern leading to this disease.