Udacity Machine Learning Engineer Nanodegree

Capstone Proposal

Ronald Mulumba
July 18, 2019

Identify Pneumothorax Disease in Chest X-rays [1]

Domain Background

A pneumothorax is an abnormal collection of air in the pleural space between the lung and the chest wall [2]. This air pushes on the outside of the lung making it collapse. A pneumothorax can be a complete lung collapse or a collapse of only a portion of the lung [3]. The symptoms of a pneumothorax typically include sudden onset of sharp, one-sided chest pain and shortness of breath. A pneumothorax can be caused by a blunt or penetrating chest injury, certain medical procedures and damage from underlying lung disease. It may also occur for no obvious reason [2][3]. In some occasions, a pneumothorax can be a life-threatening condition [3].

A pneumothorax is generally diagnosed using a chest X-ray [4]. However, they can sometimes be difficult to confirm from chest X-rays[1].

Problem Statement

The problem is a classification problem that stems from the SIIM-ACR Pneumothorax Segmentation - Identify Pneumothorax disease in chest x-rays Kaggle competition [1] where it is defined as follows:
"In this competition, you'll develop a model to classify (and if present, segment) pneumothorax from a set of chest radiographic images. If successful, you could aid in the early recognition of pneumothoraces and save lives."

Datasets and Inputs

The datatest, just like the problem stems from the SIIM-ACR Pneumothorax Segmentation - Identify Pneumothorax disease in chest x-rays Kaggle competition [1]. The data is supplied via the Cloud Healthcare API and it contains images in DICOM [5] format and annotations in

the form of image IDs and run-length-encoded (RLE) masks. Some of the images contain instances of pneumothorax (collapsed lung), which are indicated by encoded binary masks in the annotations. Images without pneumothorax have a mask value of -1.

The dataset is appropriate for the problem since it contains all the data needed to classify and segment pneumothoraces in the images.

Solution Statement

The solution will aim to predict the existence of a pneumothorax in a test image using a convolutional neural network and indicate the extent of the image using binary masks and encode them using RLE.

Benchmark Model

The benchmark model for this problem will be a simple convolutional neural network. I'll strive to improve on it's performance using other algorithms. I'll also strive to move as high as possible on the Kaggle competition leaderboard.

Evaluation Metrics

According to the Kaggle competition webpage [1], the evaluation for this competition will be the mean Dice coefficient [6].
"The Dice coefficient can be used to compare the pixel-wise agreement between a predicted segmentation and its corresponding ground truth. The formula is given by:

$$\frac{2*|X \cap Y|}{|X|+|Y|}$$

where X is the predicted set of pixels and Y is the ground truth. The Dice coefficient is defined to be 1 when both X and Y are empty. The leaderboard score is the mean of the Dice coefficients for each image in the test set" [1].
I will use the suggested Dice coefficient to measure the performance of the model.

Project Design

The first step I'll perform is to explore and visualize the dataset in order to have a good understanding of the data. The data is already split into a training and test set. However,

there is no validation set. I'll split the training set to create a validation set. The percentage of the data to use for the validation set will be decided after exploring the data and testing different percentages. The percentage that works best will be chosen.

For the benchmark model, I'll explore the possibility of using a custom convolutional neural network or use an out of the box neural network. The aim of the benchmark model will be to create a model that works better than a guess.

I'll use the transfer learning technique to train different models with different hyperparameters. The model with the best performance will be picked. I'll then create a function that creates a mask on the images that are predicted to contain pneumothorax. The model will then be tested using the test set and evaluated using the Dice coefficient method. The results will be compared to the benchmark model and the Kaggle competition leaderboard.

Abbreviations

SIIM - Society for Imaging Informatics in Medicine (SIIM)
ACR - American College of Radiology
DICOM - Digital Imaging and Communication in Medicine
API - Application Programming Interface
RLE - Run-length-encoding

References

[1] SIIM-ACR Pneumothorax Segmentation
Identify Pneumothorax disease in chest x-rays.
https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation

[2] Pneumothorax
https://en.wikipedia.org/wiki/Pneumothorax

[3] Pneumothorax
Symptoms and Causes
https://www.mayoclinic.org/diseases-conditions/pneumothorax/symptoms-causes/syc-20350367

[4] Pneumothorax
Diagnosis and Treatment
https://www.mayoclinic.org/diseases-conditions/pneumothorax/diagnosis-treatment/drc-20350372

[5] Digital Imaging and Communication in Medicine (DICOM)
https://www.dicomstandard.org/dicomweb/

[6] Sørensen–Dice coefficient
https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient