STATS507 W20 - Final Project Proposal, Group 5

Group Members: Xiaolin Cao, Yinuo Chen, Yuan Zeng

Presenter: Undetermined

Research Question:

Build an algorithm that visually detects pneumonia based on medical images.

Background:

According to WHO, pneumonia accounts for approximately 1.4 million deaths of children under 5-year-old worldwide every year. At the beginning of 2020, the novel coronavirus COVID-19 has spread across the world and caused severe damage on healthcare systems and the economy. Even under the situation where the treatment leads to full recovery, the long term consequences like pulmonary fibrosis can greatly worsen people's quality of life. Therefore, early diagnosis is critical. However, diagnosis of pneumonia consumes a lot of medical resources. It requires patients to take chest X-rays and trained radiologists to interpret the medical images. With computer-aided diagnosis of pneumonia, hospitals can speed up this process and reduce the amount of resources used.

Dataset:

Source: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data

Description:

- Number of observations:
    - Training set: 26684
    - Testing set: 3000
- Variable descriptions:
    - Both training and testing datasets comprises medical images in DICOM format.

- The training set is augmented with diagnosed labels, which contains bounding boxes of opaque areas in the lung and confidence index of the existence of pneumonia.
- The bounding box is represented by upper-left x coordinate, upper-left y coordinate, width and height. The confidence index is a binary variable (0: no evidence of pneumonia / no lung opacity, 1: has evidence of pneumonia / lung opacity).

Task:

We need to detect as many as necessary of the bounding boxes corresponding to the diagnosis of pneumonia in each chest radiography (2D high resolution grayscale medical image) to make predictions. So this is a classification problem.

Method:

- Exploratory data analysis:
  - Probe into the medical images to extract relevant information and rescale the original images to low dynamic and low bit-depth images if possible.
  - In addition to the binary classification (presence or absence of pneumonia), the bounding boxes without pneumonia are further categorized into *normal* and *no lung opacity / not normal.* This subdivision accounts for the cases in which pneumonia is excluded but other abnormalities exist. Even though the third category will not show up in our final prediction, we still include it during the model training procedure to check if its existence can improve the algorithm.
- Algorithms considered:
  - Segmentation algorithms

- Convolutional neural network
- YOLOv3
- CheXNet algorithm
- Transfer learning