Kusha Sareen

Data Selection Proposal

1. Dataset

Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, viewed 02 October 2019, <https://www.kaggle.com/nih-chest-xrays/data>.

This dataset published by the National Institute of Health contains 112,120 chest X-ray images from 30,8905 unique patients. It's expected that the labels are >90% correct which is suitable for weakly-supervised learning but some re-labelling may be required. Some scans also offer a disease region bounded box in order to better localize a significant area in the image. I chose this dataset because it is the most extensive source for frontal view chest x-rays (one of the most challenging, frequent, and costly diagnostic measures) in the public domain. The data also contains fields concerning patient information and labels 15 common pulmonary diseases.

2. Methodology

a. Preprocessing

Although I believe the dataset to be feasible, there are some challenges that come with it. In addition to standard preprocessing, we must consider how the image classification data is to be used to with the other patient characteristics/bounded box available. The data also has to be grouped into training, test and validation sets. I hope to experiment with k-fold cross validation.

b. Machine learning model

I hope to use an unsupervised learning algorithm to cluster the data and relabel outliers that were likely mislabelled. Then, I hope to use deep learning (specifically an inception network has shown prowess in classifying medical images) to predict the label of the data. This is a preliminary evaluation and I intend to modify the methodology and I continue to research the problem/visualize the data.

c. Final conceptualization

I hope to build a web app where users can upload an x-ray in addition to the required fields and get a plausible diagnosis. Specific details pertaining to the implementation of this web app will likely actualize closer to date.

Currently, the highest test accuracy for this model is around 87% (which is high considering the mislabels), however, in the past, Google was able to achieve an accuracy upwards of 95% on a similar medical imaging problem.