Discriminating COVID-19 Chest X-Rays using CNNs and Transfer Learning Across a Distributed Cluster

Allison Chilton

Laksheen Mendis

Menuka Warushavithana

Rachit Dalal

1. Problem formulation

With the global outbreak of COVID-19 (or the Coronavirus disease), one of the most vital challenges medical professionals are faced with is effectively diagnosing the disease. The problem we are focusing on is, how can one identify COVID-19 cases using chest X-Rays? At the time of this writing, testing for COVID-19 is heavily underutilized in the United States because of lack of access to test kits. However, X-Ray machines are widely available at medical labs and X-Ray images are one of the most commonly used imaging techniques used to diagnose diseases associated with lungs. We are trying to determine if it is possible to identify a high degree of correlation between a chest X-Ray and specifically COVID-19 vs other types of illnesses or healthy lungs.

This is an interesting big data problem because of the amount of hyper parameters available for Deep Learning. Especially with Convolutional Neural Networks (CNNs), there are many parameters, such as stride, kernel depth, padding, and more as well as general architecture that performing a large search of these possibilities cannot be done in a timely manner on one machine. Additionally, the datasets for these models are very large, and take many resources to process them.

As we are in the midst of the COVID-19 pandemic, the success of this experiment will be accommodating to medical professionals across the globe to quickly and efficiently identify patients who have contracted COVID-19 and proceed to treatment.

2. Strategy to solve the problem

2.1. Choice of framework:

We plan to use PyTorch as the distributed computing framework. We first considered both PyTorch and TensorFlow as they both are open-source. However, the reason why we decided to proceed with PyTorch is because of the way the two frameworks define computations graphs. Tensorflow creates a static graph, while PyTorch is operated on dynamic graphs, which means, In Tensorflow, the entire computation graph of the model has to be defined before running the machine learning model. But in PyTorch, it is possible to dynamically define the computations graph. This is particularly helpful while using variable length inputs in RNNs [3].

2.2. Datasets

Mainly, we are planning to use 2 datasets in our project.

The first dataset would be a 42 GB Chest X-Ray dataset from the National Institute of Health, which was posted 2 years ago on Kaggle [1]. This comprises 112,120 X-Ray images with disease labels from 30,805 unique patients. There are 15 classes (14 diseases, and one for "No findings"). Images can be classified as "No. findings" or one or more disease classes:

- Atelectasis
- Consolidation
- Infiltration
- Pneumothorax
- Edema
- Emphysema
- Fibrosis
- Effusion
- Pneumonia
- Pleural thickening
- Cardiomegaly
- Nodule
- Mass
- Hernia

The second dataset we will be using consists of COVID-19 cases with chest X-ray or CT images. It contains COVID-19 cases as well as MERS, SARS, and ARDS.

This particular dataset was posted on Kaggle 12 days ago [2]. The size of the dataset is 80 MB.

Although both datasets contain data beside X-Rays, during this project we will only be using X-Ray images.

2.3. Data pre-processing and explorative analytics

We will preprocess the data by normalizing them to a fixed input size determined by hyperparameter search, as well as adding grayscale filtering if it improves the result. Every image in the dataset that is not a chest X-Ray such as CT-scanned images will be filtered out. Additionally, we will seek to equalize the number of classes within the dataset as much as possible, so that the trained model does not heavily bias one prediction over others based on probability bias.

2.4. Deep Learning model application

In deep learning, Convolutional Neural Network (CNN) is a class of Deep Neural Networks, most commonly applied to analyzing visual imagery and it is mainly used in supervised learning. Transfer Learning focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. In this project, we plan to use a CNN and Transfer Learning. Below is a more informative description of our plan.

The main inputs to the neural network will be a series of convolutional layers with parameters and subsequent pooling and linearization layers determined by hyper parameter search.

We will experiment with training separate models that either work in tandem for a final output layer, or breed and evolve the model parameters using a distributed genetic algorithm.

After the first neural model is trained to sufficient level, we will begin the process of transfer learning. Because the COVID-19 dataset is much smaller, we will be using the large chest X-Ray data set to teach the model what different features of healthy and unhealthy lungs are, and then attempt to apply transfer learning techniques to the model to teach it with a diminished data set size. We will try both locking the model and adding new input and output layers, as well as leaving the model unlocked and adjusting the models gradients as a whole.

3. Evaluation Method

In this project setting, we will need to use methods to evaluate both models, i.e. model prior to transfer learning and updated model used for transfer learning. The methods we are planning to use are accuracy score, confusion matrix, precision and recall. On the other hand, we will use Area Under the Curve (AUC) Receiver Operating Characteristics (ROC) curves on the final model. More explanation about them can be found in the following subsections.

3.1. Model Prior to Transfer Learning

To evaluate the first model which is trained on Chest X-rays dataset, we will use accuracy score to figure out how many samples of the training set and testing set were correctly classified. However, we will need to be careful with getting a very high accuracy for the training set, because the model might overfit the training set. Thus, it might perform badly on the testing set. Therefore, first we need to figure out the most appropriate model which results in a reasonable training accuracy and a good testing accuracy. For instance, if the training accuracy increases and testing accuracy decreases, or if the training loss decreases, but testing loss increases while training for more epochs, the model is not performing well. Ideally, accuracy curves for both training and testing sets should increase while loss curves for both decrease, when training for more epochs.

Using the same plots (accuracy vs epochs and loss vs epochs), we will be able to identify the best number of epochs to train the Deep Neural Network.

Confusion Matrix is a method to count the number of predicted values against the true values (targets). This will create a matrix that acts as a heat map, indicating where the predicted values fall relative to the true values. Thus, as mentioned in the previous sections, the first model will be performing multi-class classification. Therefore, we will use a confusion matrix to evaluate its performance in predicting correct categories as well as where it goes wrong.

3.2. Updated model to be used for transfer learning:

Since we are using transfer learning, we will use the model trained on Chest X-rays dataset and modify it appropriately to be used with COVID-19 Chest X-Ray dataset. At this step, it is very important to evaluate how well this model works, because it was not originally trained on COVID-19 Chest X-Ray dataset.

Thus, we will use metrics such as, accuracy score, confusion matrix (along with precision and recall) on this updated model, to explore how well suited it is to use transfer learning for categorizing COVID-19 X-Rays.

Since our problem is a multi-class classification problem, we can use the AUC-ROC curve to evaluate the model's performance. AUC-ROC curve is a performance measurement for classification problems at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much a model is capable of distinguishing between classes. Higher the AUC, better the model is at distinguishing between classes. An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to 0 which means it has the worst measure of separability [4]. Thus, we will try to come up with a model with an AUC curve which is closer to 1.

4. Project Timeline (Weekly plan) & Roles / Responsibilities

Roles for each Team Member :

Rachit - Developer, Data Visualizations, Interpretation and Evaluation, Ad-hoc Analysis and Presenting results, Building Analytic System

Allison - Developer , Processing, Cleansing, and Verifying the integrity of data used for Analysis , Data Interpretation and Evaluation.

Menuka - Developer, Ad-hoc Analysis and Presenting results

Laksheen -Developer, Data Visualizations, Building Analytic System

Weekly Plan :

Task	Start Date	End Date	Status
Data Source & Data Selection	-	03/30/2020	Done
Data Cleaning	03/31/2020	04/05/2020	In-progress
Data Transformation	04/06/2020	04/12/2020	Not Started
Familiarization with technology and finalizing workflow for the analysis and interpretation. ML models learning and understanding	04/13/2020	04/19/2020	In-progress (Continuous Activity)
Analysis and Building Analytic System	04/20/2020	05/01/2020	Not Started
Interpretation & Evaluation	05/02/2020	05/07/2020	Not Started

5. Bibliography

[1] National Institutes of Health Chest, “NIH Chest X-rays,” Kaggle, 21-Feb-2018. [Online]. Available: https://www.kaggle.com/nih-chest-xrays/data. [Accessed: 31-Mar-2020].

[2] Bachir, “COVID-19 chest xray,” Kaggle, 18-Mar-2020. [Online]. Available: https://www.kaggle.com/bachrr/covid-chest-xray. [Accessed: 31-Mar-2020].

[3] U. India, “Tensorflow or PyTorch : The force is strong with which one?,” Medium, 14-Nov-2018. [Online]. Available: https://medium.com/@UdacityINDIA/tensorflow-or-pytorch-the-force-is-strong-with-which-one-68226bb7dab4. [Accessed: 31-Mar-2020].

[4] S. Narkhede, “Understanding AUC - ROC Curve,” *Medium*, 26-May-2019. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. [Accessed: 31-Mar-2020].