CSE 6250 Student Project Proposal

Hua Zhao, M.S.
Georgia Tech

Motivation

The COVID-19, as a new type of viral pneumonia, has become a pendenmic continuing to threaten the world in 2020. To fight against the desease, first step is to detect it from potential patients before proceeding to next step. One of the main early stages of screening COVID-19 is through radiography, which is either computed topography (CT) or chest X-ray (CXR) to screen patients prior to other more detailed, more costly and more time-consuming detecting methods like reverse transcriptase-polymerase chain reaction testing (PT-PCR). Compared with CT, CXR is less dangerous in amount of radiation. According to Harvard Medical School, A CXR image delivers 0.1 mSv, while a chest CT delivers 7mSv - 70 times as much. Additionally the American College of Radiology's (ACR) published guidelines in 2020-03: "CT should not be used to screen for or as a first-line test to diagnose COVID-19... CT should be used sparingly and reserved for hospitalized, symptomatic patients with specific clinical indications". These make an accurate, quick, low-cost automatic detection mechanism with CXR to help radiologists in greate need, which is also the motivation for the author to conduct this research. The ultimate goal of this reasearch, is to develope a pipeline with current big data systems and other data management tool to more conveniently train and fine-tune introduced state-of-art deeping learning algorithms, in the hope of both creating a larger annnotated dataset for open-source and building a more accurate algorithm compared to prior researches. The actual direction of the reasearch should depend on how the research actually goes before the submission deadline of 2020-12.

Literature Survey

Detecting COVID is a binary or multi-class classification problem. In this research, we'll have 2 tasks for models: 1. classify as COVID vs. non-COVID. 2. classify as COVID vs. other pneumonia vs. none neumonia.

Since the growth of the deeping learning area, scientific community has focused on development of artificial intelligence (AI) algorithms to help detect pneumonia patients. After the COVID-19 outbreak, the area is even heated. In the CXR imaging area, convolutionary neural networks (CNNS) have been use the most. For example, Tartaglione *et al.*[1] in a study applied transfer learning algorithsm - both ResNet-18 and ResNet-50. They also used a pre-trained encoder to pre-extract image features prior to CNNs; Wang *et al.*[2] built a COVID-net - a derivative of CNNs tailored to detect COVID-19 from CXR with more than 400 chest X-ray images and more than 10,000 total images and achieved good accuracy.

However, there are still some issues we should be carefully dealing with. First, it is the limited amount of COVID-19 images. Based on the limit of public data, many COVID-19 CXR papers trained their models with less than 500 COVID-19 positive images. Even when Wang *et al.*[2] merged 5 datasets from publications and collections and managed to have 400+ COVID images with total images over 19,00 in 2020-05. It's still very limited given the needs of large and balanced dataset to well-train a complex network to obtain promising classification ability. This makes the reasearch both obtain more authuritative dataset, or be more careful with data augumentation when dealing with input dataset. (This will be detailed as the researach goes.). As for obtaining more COVID-19 datasets, Vaya *et al.*[3] published a new comprehensive dataset including 1,380 CX and 885 DX which may be helpful to this research.

The second issue is CNNs may extract features irrelevant to medical information in CXR images, making the network classifies images with good resutls based on non-medical information from iamges. Maguolo *et al.*[4] in a study trained AlexNet with different sources of COVID CXR datasets where the center images part (mainly the body) were turned to balck, and the network was able to indentify the specific source of dataset with a suprisingly high accuracy. When Tartaglione *et al.*[1] in their study were trainign ResNet-18 and ResNet-50, similar things happend such that when the same source of dataset is used in training and testing, the AUC and accurarcy tend to be very high while when training dataset and testing dataset is from different source, those metrics decline significantly - some of the AUCs are even less than 0.50. However, as Bernheim *et al.*[5] pointed out in a medical study, the main features in COVID CXR images are ground-glass opacity (GGO) and consolidation. This makes training a network with medical-relevant features in

CXR essentially important for the purpose of this research.

Data

We'll be merging 7 datasets in this research, and hope to gain more as the research goes. Current total collected CXR images (below) will be as follows:

	COVID	other pneumonia	no-finding
Count	3725	74321	64239

How the data will be processed will be as follows:

1. Datasets preprocessing stage 1: merge datasets from various sources, carefully handle overlaps
2. Datasets preprocessing stage 2: Histogram Equialization as a mean to reduce contract different among images and try to guarantee uniform image dynamics across images.
3. Datasets preprocessing stage 3: Data segmentation of lung area. Train a U-net (Ronneberger *et al.*[6]) to aviod medical-irrelavent information outside of the lung in images. Sharp edges will be blurred by 3-pixel radius.

In addition, a validation group will be formed by randomly selecting a set of images from the total which the center of image will be turned to black as comparison group to evaluation final model performance.

All images used will be AP or PA, no lateral images will be used for this research.

The sources are as follows:

- A COVID-19 dataset collection project by Cohen *et al.*
- A COVID-19 dataset collection project by Wang *et al.*
- Another COVID-19 dataset collection project by Wang *et al.*
- COVID dataset from kaggle: COVID-19 Radiography Database
- non-COVID dataset from kaggle: RSNA Pneumonia Detection Challenge
- large COVID dataset by Vaya *et al.*[3]
- large non-COVID dataset from chexpert

Approach

For the algorithm, the author plans to explore some convolutionary neural networks, especially with transfer learning from pre-trained ResNet, and COVID-net from Wang *et al.*[2]. The procedure of the model will be as follows:

1. Datasets preprocessing as defined in **Data**
2. Pre-training stage: pre-train on the feature extractor on a large non-COVID CXR datasets from chexpert.
3. Training detail 1: fine-tuning feature extractor on COVID datasets
4. Training detail 2: Data augmentation to balance COVID-dataset: translation, rotation, horizontal flip, zoom, and intensity shift, etc,. Detail implementation may vary as research goes.
5. Testing stage: Two groups: One group is testing group which will be normal images to test. The other group is validation group, whoses image center will be turned to black as comparison.
6. Evaluation stage: Evaluation model performance based on four metrics: **AUC, Accuracy, Balanced Accuracy, Diagnostic odds ratio (DOR)**

Experimental Setup

The computing envioronment of the research is planned as follows:

```
Computing Environment of COVID-19 Detection Research
|-- Win 10.0.18362 N/A Build 18362 (localhost)
|-- Spark (kernal level pipelining in ETL and algorithm computing)
|-- Python
|   |-- pyspark (API to spark)
|   |-- pandas
|   |-- scipy
|   |-- dvc (data level pipelining in whole model)
|   |-- tensorflow
|   |-- pytorch
```

Note that actual environment may vary in details as research goes.

Timeline

Since there's only author of the research project, all tasks will be distributed to the only author.

1. 2020-10-18: Finish Datasets preprocessing stage 1: merge datasets from various sources
2. 2020-10-25: Finish Datasets preprosses through stage 2 and 3: Histogram Equialization and lung segmentation.
3. 2020-11-01: Finish pre-training stage and first iteration of training, testing
4. 2020-11-11: Finish another iterations of training and testing, draft research with resutls and/or new findings
5. 2020-11-25: Finish refining of the model process, fine-tune model parameters with new model iterations
6. 2020-12-06: present

References

1. Enzo Tartaglione, Carlo Alberto Barbano, Claudio Berzovini, Marco Calandri, Marco Grangetto, Unveiling COVID-19 from Chest X-ray with deep learning: a hurdles race with small data, *arXiv:2004.05405*, 2020-04
2. Linda Wang, Alexander Wong, COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images, *arXiv:2003.09871*, 2020-05
3. Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, Marisa Caparrós, Germán González, Jose María Salinas, BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients, *arXiv:2006.01174v3*, 2020-06
4. Gianluca Maguolo, Loris Nanni, A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images, *arXiv:2004.12823*, 2020-09
5. Adam Bernheim, Xueyan Mei, Mingqian Huang, Yang Yang, Zahi A. Fayad, Ning Zhang, Kaiyue Diao, Bin Lin, Xiqi Zhu, Kunwei Li, Shaolin Li, Hong Shan, Adam Jacobi, Michael Chung, Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection, *RSNA Radioalogy, https://doi.org/10.1148/radiol.2020200463*, 2020
6. Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, *arXiv:1505.04597*, 2015