Classifying Covid-19 patients based on lung x-rays

Kevin Lee and Rakshitha Kori Raj

Abstract

Image classifier networks are a powerful tool for medical professionals to assist in diagnosing diseases or other illnesses. In the case of Covid-19, we have a very new disease that has quickly become an issue of importance due to the high infection rate. By training a classifier to determine whether an x-ray image of a patient's lungs is indicative of Covid-19 instead of just some other respiratory disease or a healthy patient, we can assist medical professionals in their task to limit the spread of the highly infectious disease. Adding the ability to x-ray a patient to determine whether or not they have Covid-19 would be a solid alternative to the current tests which are in high demand and limited supply.

Project Schedule (Subtasks, approach, deadlines, including research)

Literature Research:

1. Segments of Lungs x-ray Apr 1 - Apr 4 Rakshitha
2. Work that has already been done on deep neural networks done on x-ray images, especially Pneumonia/SARS detection. Apr 1 - Apr 4 Kevin

Tasks:

1. Image preprocessing image extraction Apr 6 Apr 13 Kevin/Rakshitha (?)
    a. Normalizing contrast
    b. Resizing and reshaping images
    c. Removing artifacts
2. Dataset augmentation Kevin Apr 6 - Apr 13
3. Initial CNN Rakshitha Apr 6 - Apr 13
4. Generate Visualizations Saliency Maps and grad-CAMs Kevin Apr 13 - Apr 20
5. Iterate steps 2 and 3 to fine tune parameters and come up with better results Apr 20 - Apr 27 Kevin/Rakshitha
6. Final Project report start week before final date Rakshitha
7. Preparation of slides for final presentation week before final date Kevin
8. Try to get other predictions like if the patient survived , duration for which the patient could have had the virus. Kevin/Rakshitha further analysis.

Dataset augmentation

There are a considerable number of hurdles we must overcome before being able to successfully create a classifier. First, the dataset(s) we are using are fairly limited in scope, as the disease is incredibly recent and much manpower is being spent on mitigating the spread and assisting those who most need it rather than doing x-rays or other non-standard screenings. Covid-19 attacks the respiratory system, so there are other ways to determine if a patient has it than doing an x-ray. Since it isn't a common procedure for testing for Covid-19, more or larger datasets will be hard to find. In order to have a reasonable number of input images for the classifier, we will need to augment the dataset(s) in some way. While we were suggested a data

augmenting package[2], we would also be able to use similar techniques to augment the datasets ourselves.

Alongside the issue of the small dataset is that of classification. After we initially viewed the paired images of healthy and infected individuals, it was difficult for us to view any kind of difference in the x-rays. After speaking to a professional, we learned that the x-rays of individuals with Covid-19 have lungs with fuzzy peripherals. The lungs should look somewhat sandy, glassy looking, or patchy. While informative, this information is somewhat difficult to pick up as a layman, so the classifier may have difficulty in that the differences are highly subtle.

Image preprocessing

Additionally, there are a considerable number of hurdles that must be overcome in the image pre-processing stage. This is likely where most of the work will have to be done. The issue with using such a recent dataset is that we are highly limited for options. A preliminary glance at the only available datasets reveals that there is a significant amount of diversity in both healthy and infected patients. There are clearly young, old, thin, and overweight individuals, and this contributes to size and shape differences. It may be necessary to scale and shear the images to get them in similar orientations and scales. Features like the corner of the right (left side of image) diaphragm and the spine could be used as localizing features for alignment. The images also contain a lot of difference in lighting and contrast, although this is simple to fix in comparison.

Unfortunately, there are some issues that may be difficult or impossible to overcome. It is clear that some of the patients in the dataset had additional health issues other than Covid-19, so the data is not as clean as it could be. Covid-19 also causes more symptoms in those with ongoing chronic diseases, so it is likely our dataset contains a disproportionate number of individuals who have other health issues. This could create some confounding information for the classifier. Since our best identifier for individuals with Covid-19 is fuzziness in the lung peripherals, we must be very careful with our de-noising. Too much blurring could result in images that are too similar in sharpness (or lack thereof), and would make it even more difficult for the classifier. There are also some artifacts in the images that are difficult to remove (ecg electrodes, wires, text labels) without manually editing or overly strong blurring.

Train binary classifier

After pre-processing, we can train a deep convolutional neural network to do binary classification. With the augmented dataset, we can do k-fold cross validation to create training and validation sets. From there, all that needs to be done is to train the network and iterate by tuning hyperparameters and methodology.

Visualize CNN to verify that it makes sense

Ziming recommended “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps” for a visual representation of how a convolutional neural network weights certain areas of the image. This would be a great tool to use at the end of the project to see if our CNN is focusing on the lung areas. These Visualization techniques to test and adjust parameters are very critical to this project as we do not have a large data set[4].

References

[1] Joseph Paul Cohen and Paul Morrison and Lan Dao COVID-19 image data collection. Accessed on: Mar. 31, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[2] Nabeel Sajid Covid dataset and augmentation method. Accessed on: Mar. 31, 2020. [Online]. Available: https://www.kaggle.com/nabeelsajid917/covid-19-x-ray-10000-images

[3] Zhe Xu MD et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Accessed on: Mar. 31, 2020. [Online]. Available: https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(20)30076-X/fulltext

[4] Pasa, F., Golkov, V., Pfeiffer, F. et al. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. Sci Rep 9, 6268 (2019). Accessed on: Mar. 31, 2020. [Online]. Available: https://doi.org/10.1038/s41598-019-42557-4

[5] Chest X-ray (CXR) Interpretation https://geekymedics.com/chest-x-ray-interpretation-a-methodical-approach/