CS598 DL4H Spring 2022 Project Draft

Ashutosh Agarwal and Chandan Goel
aa61@illinois.edu, chandan3@illinois.edu

Group ID: X (TBD), Paper ID: External (link TBD)
Presentation link (TODO): https://www.youtube.com
Code link: https://github.com/AshutoshAgarwal01/CovidPred_Repro

1 Introduction

This work is done to build efficient deep leaning models using the convolutional neural network (CNN) using a very limited set of chest X-ray images to differentiate COVID-19 cases from healthy cases and other types of illnesses. The overall goal of the paper is to train models such that rapid screening of COVID-19 patients is possible in a non-invasive and automated fashion.

Several other studies were done before this work to classify COVID-19 patients using chest X-ray images. This study recognizes following issues in classification of COVID-19 based on chest X-ray that result in poor performance of models.

- Scarcity of x-ray images available to train a viable deep learning model.
- Biased learning of the deep learning based model when images of multiple age groups are combined together to form a data set. E.g., x-ray images of pediatric patients combined with adult patients.

In this study, authors have applied following techniques to overcome issues mentioned above:

- To overcome scarcity of x-ray images, authors proposed creating a much larger artificial dataset using smaller original dataset. This artificial dataset was created by applying multiple image augmentation techniques on the original data.
- Careful image selection: Authors propose to use similar type of images (same view, same age group) to train a very targeted model. This study claims that models built with similar type of images perform better than those models that take a wide variety of images.

Overall, authors showed that artificially generated x-ray images using image augmentation techniques greatly improved model performance when compared with original smaller set of images.

2 Scope of reproducibility

Due to data scarcity related to COVID-19 chest X-ray images, there was very small set of data available for this work. To overcome this issue, authors artificially generated large number of images using 25 augmentation techniques on original images and used this data for model training purposes.

"This study shows that a deep learning model trained with larger volume of artificially generated X-ray images using image augmentation techniques will have higher accuracy than the model trained with small set of original images."

Data scarcity is a well-known problem in the field of healthcare analytics, especially when research is aimed at fairly new area e.g., classification of COVID-19 patients when the pandemic just broke out. We are motivated to take on this study as it tries to solve a real life problem which exists across all healthcare domain. It will not only help improve model performance, but also reduce cost and time involved in getting sufficient size data for any healthcare study.

2.1 Addressed claims from the original paper

We will be testing following two claims from the paper.

- **Claim 1**: Model trained with original and augmented images combined results in higher accuracy across all classification labels.

  Due to data scarcity related to COVID-19 chest X-ray images, there was very small set of data available for this work. It makes model training and validation very difficult. Performance of model was not good. To overcome

this issue, authors artificially generated large number of images using 25 augmentation techniques on original images.

Further, authors compared performance of following two models using an unseen external data.

- Model with only original images.
- Model with original images along with augmented images.

Authors concluded that model trained with original and augmented images combined, results in higher accuracy (True positive rate) across classification labels. In our project work, we want to confirm this claim.

- **Claim 2**: Models trained with 120 and 140 degree rotated images results in higher accuracy (True positive rate) across classification labels when compared with model trained with original images.

  Authors showed that models trained with 120 and 140 degree rotated images were complimentary to each other. i.e., if one model performed bad for certain label than other model performed better for the same label. Together, these two models beat performance of model trained with original images.

3 Methodology

The authors of the original paper made a diligent effort for anyone to understand their paper and code easily. They provided link to their GitHub repository CovidPred [4] where code was present and they clearly mentioned source of data that they used for their study.

We are planning to write new code for data processing and re-using (with moderate modifications) existing code for model training and testing. Following sections provide more details about this.

3.1 Model descriptions

TODO: Describe the models used in the original paper, including the architecture, learning objective and the number of parameters.

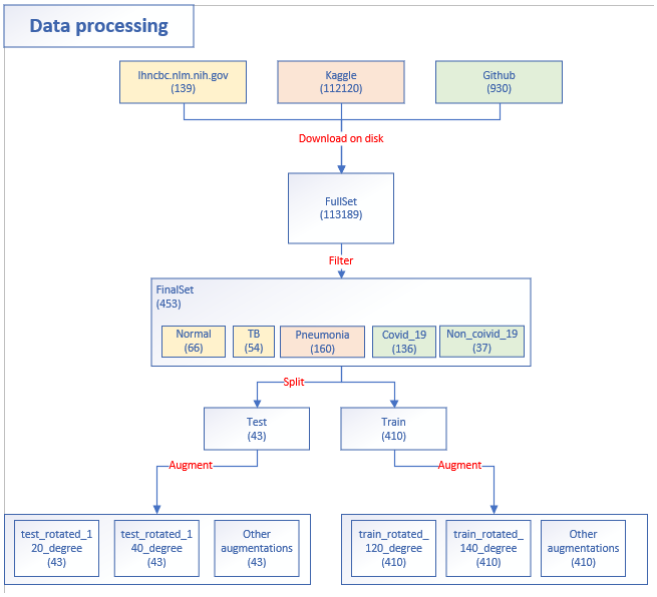
3.2 Data descriptions

Data from following three sources is used in this paper.

- github (Cohen's covid-chestxray-dataset) [1]: This dataset is used to get 'covid 19' and 'non covid 19' images
- Kaggle NIH dataset [2]: This dataset is used to get Pneumonia images.
- National Library of medicine [3]: This dataset is used to get normal and TB images.

Following diagram depicts all steps we performed to gather and process data for model training and validation purposes.

TODO: describe steps



Note: Number in each box represents number of images.

We filtered images from these datasets using same filtering criteria cited by authors in their paper. Following is the general filtering criteria.

- Age of patient must be 19 years or older.
- Only chest X-ray images.
- X-ray image view must be PA.

After filtering the data, we divided the dataset into two parts using random sampling. We reserved 10 percent of the data for external validation (test set) and remaining 90 percent for model training (training set).

After this, we created 25 new datasets by applying different augmentation techniques on original set of training and test images. Some of the augmentation techniques used are:

- Rotate images by 45, 60, 90, 120, 140 and 160 degrees

- Raise blue, green, red and hue
- Crop images
- Flip images horizontally, vertically and in both directions.
- Introduce blur to the images.

We used CloDSA [5] library for image augmentation.

We further created one more dataset by combining original dataset and all augmented datasets. Thus we had total 27 datasets. Following table shows distribution of images.

Label	Total count of images	Selected after filtering	Original dataset		Each augmented dataset (25)		Combined dataset	
			Train	Test	Train	Test	Train	Test
COVID-19	930	136	123	13	123	13	3198	338
Non-COVID-19		37	34	3	34	3	884	78
Pneumonia	112120	160	144	16	144	16	3744	416
TB (Montgomery Country X-ray Set)	139	54	49	5	49	5	1274	130
Normal (Montgomery Country X-ray Set)		66	60	6	60	6	1560	156
Total	**113189**	**453**	**410**	**43**	**410**	**43**	**10660**	**1118**
			453		**453**		**11778**	

3.3 Hyperparameters

For initial reproduction work, we used same hyperparameters used in the paper except number of iterations. Following table describes the hyperparameters and their values across different datasets.

Green boxes represent hyperparameters that were modified for our study. We had to modify these hyperparameters since number of images in our study and original study was different.

Original image/ single augmentation based models (120 and 140 degree rotation)		
Layer	**Parameter**	**Value**
	Number of iterations	529
	Batch size	16
	Number of epochs	24
	Image size	256
	Internal validation size	10%
Convolution layer 1	Filter (Kernal) Size	3
Convolution layer 1	Number of filters (out channels)	32
Convolution layer 2	Filter (Kernal) Size	5
Convolution layer 2	Number of filters (out channels	64
Convolution layer 3	Filter (Kernal) Size	7
Convolution layer 3	Number of filters (out channels	128
Fully connected layer 1	output size	256
Fully connected layer 2	output size	5

Combined augmentation based model	
Parameter	**Value**
Number of iterations	6624
Batch size	32
Number of epochs	24
Other parameters	Same as original/ single augmentation based models

3.4 Implementation

We thoroughly studied the code available in author's git repository [4]. Following is distribution of code that we wrote and re-used.

- Data processing: We wrote our own code from scratch for data processing.
- Model training and validation: We made some improvements in existing code like logging, refactoring, minor performance improvements and training/ validating models using multiple datasets by one script etc. But overall, it is heavily inspired by original work.

Code written for reproduction work can be found here: https://github.com/AshutoshAgarwal01/CovidPred_Repro

3.5 Computational requirements

In the proposal we mentioned that we will use GPU if training model with large dataset (all augmented images combined) takes very long time. However in our initial runs, even the largest dataset completed in reasonable amount of time. Therefore we decided to use the alternate in-premise CPU based hardware. Following table (Table 1) shows configuration of desktop that we used.

Processor	Intel(R) Xeon(R) CPU E5-1650 v4 @ 3.60GHz 3.60 GHz
Number of cores	6
Memory	32 GB
OS	Windows 10 Enterprise

Table 1: Machine configuration

Table 2 details summary of CPU time, disk space and memory we consumed for completing one execution of experiment. This includes data processing, model training and testing.

Total disk space needed	121 GB
Total CPU time	9 hours
Max memory requirement exclusively for the CPU process running the code.	8 GB
Suggested minimum machine memory (including OS and other processes) to run this experiment.	32 GB

Table 2: Summary of resources

Following table summarizes total time and average time per epoch for each model.

Model training and validation				
	Model size on disk (MB)	Memory consumption (GB)	** Training time	Average training time per epoch (24 epochs)
Model with 120 degree rotated images	389	2	24 minutes	1 minute
Model with 140 degree rotated images	389	2	24 minutes	2 minute
Model with original images	389	2	24 minutes	3 minute
Model with original images and all augmented images	389	8	5 hours 52 minutes	15 minutes
	1556	14	7 hour 41 minutes	

Total size of original, filtered, augmented and combined data together was **119 GB** on disk.

4 Results

Our reproduction study did not result in exact same accuracy numbers as the original study. However overall trend of the accuracy (true positive) per label is consistent with the original study.

The results obtained by our reproduction study support both the claims cited in section 2 of this paper. We will describe them in following sections.

4.1 Result 1

The paper claimed that model trained with original and augmented images combined results in higher accuracy (true positive) across all classification labels when compared with model trained with original images only.

Following table summarizes true positive rate (proportions of images correctly classified by the model for given label) of both models on unseen original data.

We can see that model trained with combined dataset outperforms model trained with original images only with large difference. This upholds the paper's conclusion that it performs much better than the baseline.

Comparing performance of model trained with original images with model trained with all images combined (original + 25 augmented).		
Label	Original Images	Combined (Tested against unseed original dataset)
Normal	50	100
Covid-19	53.85	76.92
Non-covid-19	66.67	33.33
Pneumonia	75	81.25
Tuberculosis	80	100

4.2 Result 2

Authors showed that models trained with 120 and 140 degree rotated images were complimentary to each other. i.e., if one model performed bad for certain label than other model performed better for the same label. Together, these two models beat performance of model trained with original images.

Following table summarizes true positive rate (proportions of images correctly classified by the model for given label) for all three models on unseen original data.

We can see that for all labels except Tuberculosis authors' conclusion holds good. However, both augmented models perform worse than baseline for Tuberculosis. Therefore, this claim is only partially supported by our reproduction study.

Comparing performance of model trained with original images with model trained with augmented images.			
		Augmentation (tested against augmented test dataset)	
Label	Original Images	Rotate 120 degree	Rotate 140 degree
Normal	50	100	100
Covid-19	53.85	84.62	23.08
Non-covid-19	66.67	66.67	100
Pneumonia	75	81.25	87.5
Tuberculosis	80	20	40

4.3 Additional results not present in the original paper

TODO

5 Discussion

TODO

5.1 What was easy

TODO

5.2 What was difficult

TODO

5.3 Recommendations for reproducibility

TODO

6 Communication with original authors

TODO

References

1. Cohen's covid-chestxray-dataset: https://github.com/ieee8023/covid-chestxray-dataset
2. Kaggle NIH dataset: https://www.kaggle.com/datasets/nih-chest-xrays/data
3. National Library of medicine: https://lhncbc.nlm.nih.gov/LHC-publications/pubs/

TuberculosisChestXrayImageDataSets.html

4. CovidPred: https://github.com/arunsharma8osdd/covidpred
5. CloDSA: https://github.com/joheras/CLoDSA