COVID X-Ray image classification using CRNN and COVID-Net

Hua Zhao
Georgia Institute of Technology

Abstract

The COVID-19, as a new type of viral pneumonia, has become a pandemic continuing to threaten the world in 2020. To fight against the disease, first step is to screen it under an easy-carried, fast-implemented, low-cost low-danger way typically Chest X-Ray (CXR). Nowadays more and more research has been exploring and utilizing state of art artificial intelligence algorithms, some of them being quite successful in terms of performance. In this research, the author explored some state of art neural networks built by previous researchers, including transfer learning from RestNet[2] *and COVID-net*[3]*, with a built automatic pipeline, and evaluated their performances. The research collected 22,049 images in use, which is by far, according to the best of authors knowledge, the largest dataset used in the COVID-19 CXR classification problem. Through two iterations of model training, and under limited local machines computing capability, the final COVID-Net in iteration 1 obtains (0.77, 0.68, 0.88) sensitivity and (0.35, 0.76, 0.89) precision for COVID-19/normal/pneumonia respectively; the final ResNet-50 in iteration 2 obtains (0.85, 0.97, 0.81) sensitivity and (0.54, 0.85, 0.97) precision correspondingly. Meanwhile, with Grad-CAM visualization, we can see plotted areas of recognized features from our trained ResNet-50, showing major features are located reasonably inside lung areas. Finally, the author demonstrated a complete automated pipeline that utilizes DVC through an acyclic directed graph to version and reproduce the whole procedure from Spark ETL, pytorch/tensorflow model setup, to training, plotting, testing. The architecture can be utilized for future training with your customizations on specific modules in the pipeline. See code repo in GitHub*[1]*.*

Keywords: Chest X-ray, COVID-19, AI, classification, COVID-Net, ResNet

Introduction

The research aims to train deep neural networks it set up, based on collected public dataset, to classify normal/COVID-19/pneumonia cases from CXR images, and try to achieve good and meaningful results. Evaluation metrics are confusion matrix, and TPR/Sensitivity/Recall of each label, PPV/Recall of each label, for pytorch models, also loss learning curves as function of epochs. Target model architectures include Linda Wang et els COVID-Net[3] and Kaiming He et el's ResNet[2, 10]. Some training details or processes differ from previous study in following ways:

- Adaptive/global historgram equilalization technique to increase contrast, making features more obvious
- Data segmentation techniqu to make model focus on searching features within lung areas in CXR.
- A large dataset of 22,049 CXR images are used in training, validating.
- Spark is used as PySpark package with Python to do ETL for images collection.
- A DVC pipelining technique manages the dataflow and model reproduction, with a parameter yaml file as input to control all the hyper- and non-hyper parameters used in the production. It utilizes dependency graphs and MD5 checksum to manage data version control, with data serialization in caching.

Problems

However, there have been some challenges in utilizing neural network algorithms to help with COVID-19 CXR classification. First, it is the limited number of COVID-19 images publically available to train AI. Due to the limit of public data, many COVID-19 CXR papers trained their models with less than 500 COVID-19 positive images. The largest collection prior to the paper used in the COVID-19 CXR classification problem is Linda Wang, et al[3] which contains 400+ COVID images with total images of 600+ in 2020-05. The limited datasets cultivate 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. As for obtaining more COVID-19 datasets, Vaya et al.[4] published a new comprehensive dataset including 1,380 CX and 885 DX, which are partly used in this research data collection.

The second issue is neural networks may extract features irrelevant to medical information in CXR images, making the network classifies images with good results based on non-medical information from images. Magnolol et al[5] in a study trained AlexNet with different sources of COVID CXR datasets where the center images part (mainly the body) were turned to black, and the network was able to identify the specific source of dataset with a surprisingly high accuracy. When Tartaglione et al[6] in their study were training ResNet-18 and ResNet-50, similar things happened when the same source of dataset is used in training and testing - the AUC and accuracy tend to be very high when training dataset and testing dataset are from same source; when training dataset and testing dataset are from different source, those metrics decline significantly - some of the AUCs are even less than 0.50. However, as Bernheim et al[7] pointed out in a medical study: the hallmarks of COVID-19 infection on images were bilateral and peripheral ground-glass and consolidative pulmonary opacities which resides obviously within lung areas. This makes training a network with medical-relevant features in CXR essentially important for the purpose of this research.

The third potential probelm we may pay attention to is in many early phase of COVID-19 cases, topography studies (CT, CXR) show same patterns as normal cases to radiologists. According to Bernheim et al[7], Some characteristic chest CT imaging features of coronavirus disease 2019 are related to time course of infection, with certain findings occurring with increased frequency as the time between symptom onset and initial chest CT lengthens. In a case study of 36 COVID-19 patients, 20 of the 36 patients (0.56) imaged in the early phase had a normal CT scan. Bilateral lung involvement was observed in 10 of the 36 early patients (0.28), 25 of the 33 intermediate patients (0.76), and 22 of the 25 late patients (0.88). That said, if a dataset contains different stages of CXR for COVID-19 par of samples, it will be normal to have a not high precision and/or recall due to non-obvious medical features. CT imaging is so, let alone the CXR. If a model claims over 0.90 precision or recall on COVID-19 CXR, then it may be too good to be true.

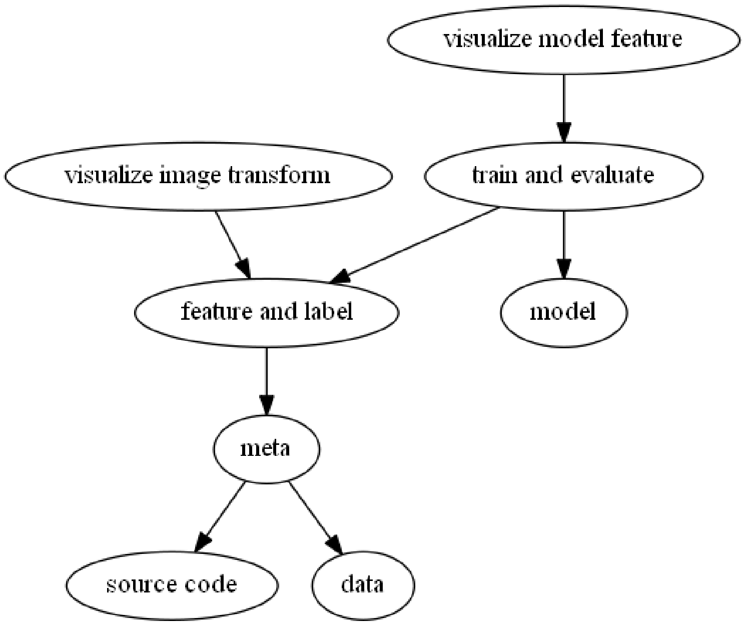

Figure 1: Bernheim et al[7]: Bar graph shows frequency of CT findings as function of time

Pipeline

The whole project process is managed by a automatic pipeline built on DVC (Data Version Control). DVC caches and versions data flow, constructs a DAG (directed acyclic graph) used to reproduce the whole procedure. The DAG consists of series of ordered stages with dependenceis and outputs including hyperparameter setting. Each stage executes an OS-dependant command, thus the whole pipeline executes a series of ordered and versioned stages through commands, from processing meta data from source images, preprocessing images, loading input hyperparameters, loading model, training, evaluating, saving performance metrics and trained model with checkpoints.

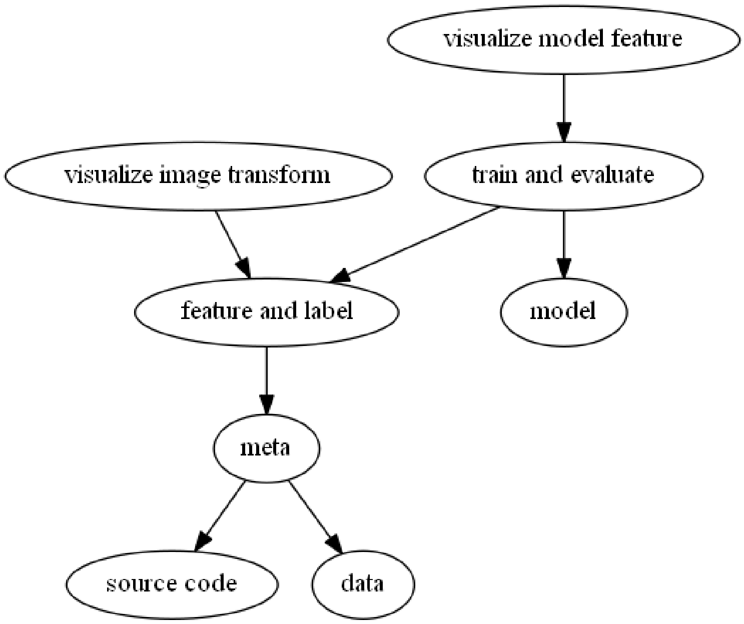

Figure 2: DAG (directed asyclic graph)

Data

CXR image and label data are collected from four public sources. See GitHub Repo ReadME[8] for data sources. Case counts are as follows:

case	pneumonia	normal	COVID-19
total	11092	10340	617
train	8873	8272	493
test	2219	2068	124

As you can see, although we have over 20,000 images, target COVID-19 label is very unbalanced. To balance samples during batch genearting for model training, batch is generated from cases of only pneumonia and normal, then we bootstrap a siginificant weight of COVID-19 cases from its set and randomly replace original generated batch to obtained a label-balanced batch for training. Here the weight is a hyperparameter into the pipeline.

Method and Implementation: ETL, Image Transformation

To process source data into cached datasets ready for model training, the process goes through two relavent stages: 1. process meta data: processs and create a meta data table with unique image id, file name, source id, and other neecssary information. The process is done through Spark, implemented as PySpark API with Python. 2: process image data: process images into disired forms. It can be fine-tuned from hyperparameters. During the second stage, image tranformation is applied, In 1st iteration, in addition to simply resizing to target size with crops, image transformation includes adaptive histogram equilization (CLAHE) implemented in OpenCV is applied, with tilesize = 8*8, cliplimit = 2.0.

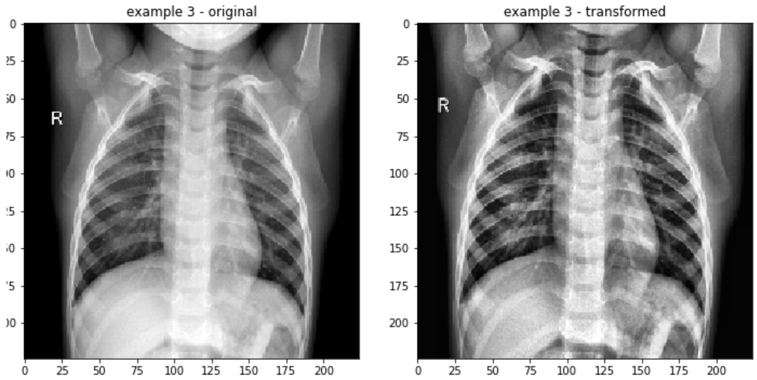

Figure 3: Example of iteration 1: adaptive histogram transformation

In 2nd iteration, based on first one, image is also transformed through lung segmentation implemented in OpenCV.

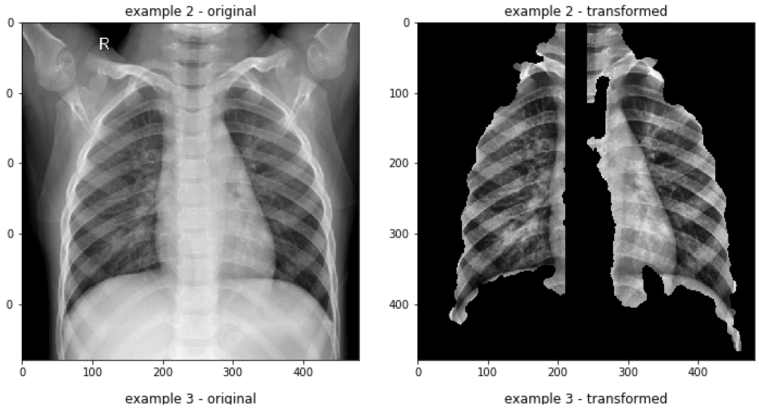

Figure 4: Example of iteration 2: segmentaion + CLAHE

Method and Implementation: Model Training

The author trained three models: ResNet-18, ResNet-50, COVID-Net. ResNets are very deep residual convolutional neural works proposed by Kaiming He et al[2]. It has resdidual representation and shorcut connection structure with heavy use of batch normalization in each layer, since 2016, ResNets has been widely used by computer vision researchers because of its depth and performance; COVID-Net is comparatively a new architect, proposed by Linda et al[3] that is a human-machine collaborative design strategy consisting of stage 1: human-driven design netowrk principles and stage 2: machine-drive design exploration. It shows a at least same accuracy given lower computational cose compared to ResNet and VGGNet, according to Wang et al[3].

Since ResNet and COVID-Net are implemented on different tools: pytorch and tensorflow, the pipeline includes both models. During training, augmentation (random horizontal flip, random rotation, random gray sacle, etc.) is applied to each image in batch, to make models less sensitive to picture pattern itself.

Main metrics used to measure and compare performance of trained models are sensitivity, precision from confusion matrix for each label. For ResNet, it also looks at loss learning curves through epochs.

Experimental Results

The author completed two ierations of the process. Due to computation power limit of author's local machine and large number of training data, completing one epoch of 20,000+ images of training and evluating usually takes 4 to 5 hours. As a result, in each iteration, models are only trained within a few epochs, however they already showed a local optimal weights of feature tensors since in each epoch, there're more than 20,000 images. The iterations are described as follows:

iteration	1	2
image shape	224, 224, 3	480, 480, 1
adaptive histogram equilization	not used	used
segmentation	not used	used
COVID-Net	trained	not trained
ResNet-50	trained	not trained
ResNet-18	trained	trained

For COVID-Net, the learning curve of TPR/sensitivity and PPV/precision are as follows:

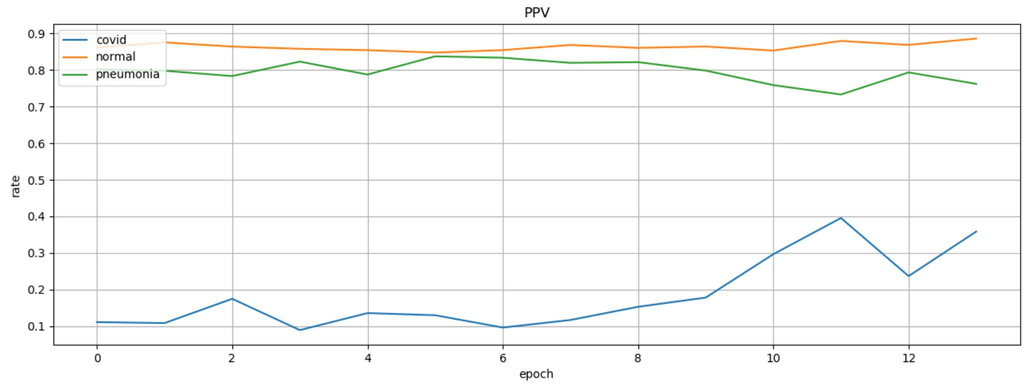

Figure 5: learning curve: PPV/precision

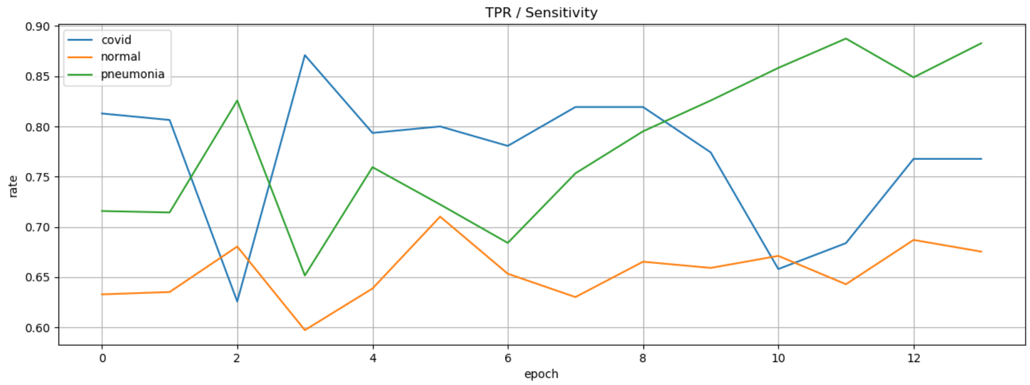

Figure 6: learning curve: TPR/Sensitivity

As we can see, the model with highest COVID-19 TPR/sensitivity is 0.87 at 4th epoch; it normalized confusion matrix displayed below:

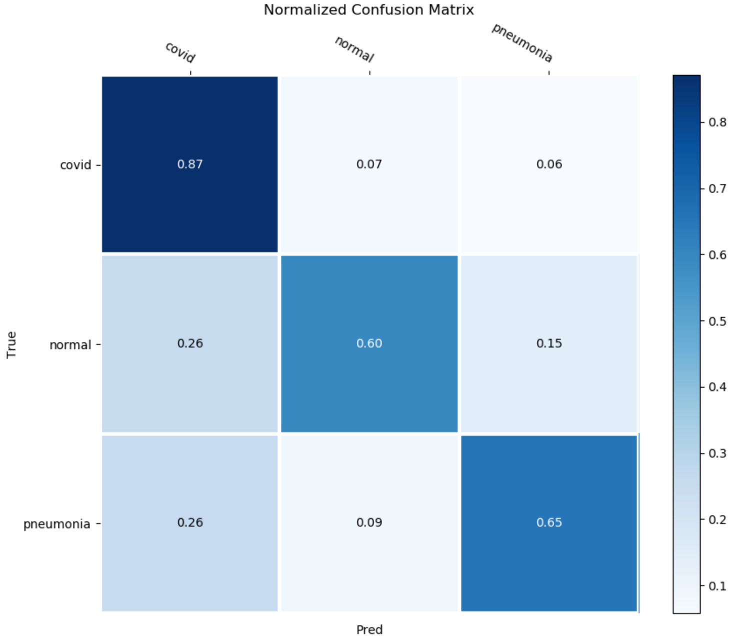

Figure 7: norm confusion matrix at 4th epoch / highest COVID-19 TPR

However this sacrificed the sensitivity for label pneumonia and label normal, as well as precision for label COVID-19 (0.18). At the final epoch, the confusion matrix is shown as follow:

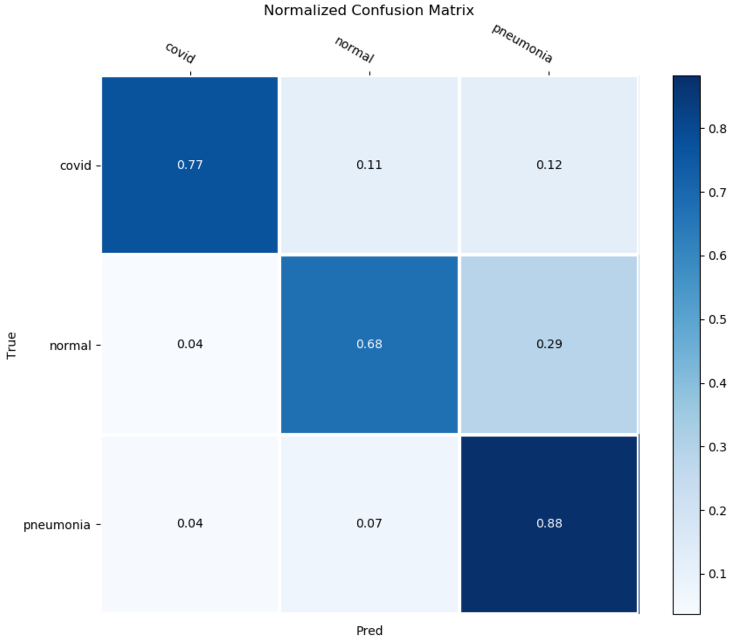

Figure 8: norm confusion matrix at 14th epoch

From the final epoch, with 0.10 decrease in COVID-19 sensitivity, the model gained 0.08 normal sensitivity, 0.23 pneumonia sensitivity, around 0.20 COVID-19 precision.

Adding COVID-Net and ResNet, together, we can write a table to describe their performance using their common metrics:

metric	COVID-19 sensitivity	COVID-19 precision
COVID-Net, iter 1	0.77	0.38
ResNet-50, iter 1	0.98	0.12
ResNet-18, iter 1	1.00	0.07
ResNet-18, iter 2	0.85	0.55

Obviously ResNet-18 in iteration 2 achived the best performance among all. It has a not bad sensitivity for COVID-19, not too high as 0.98 or 1.00 as compared to the other 2 which probably indicates overfitting as the reason discussed in Problems section Point 4. It also has been trained to gain a not too bad precision: 0.55. Taking a look at its two normalized confusion matrix:

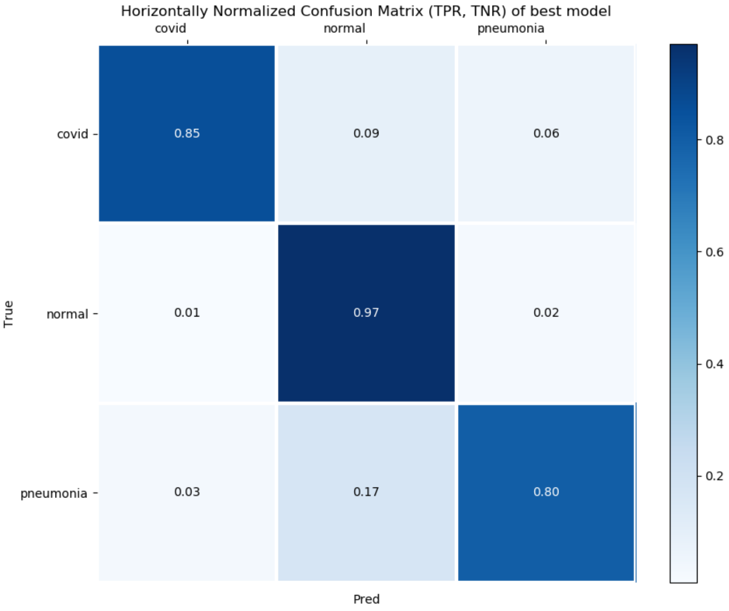

Figure 9: hnorm confusion matrix 1

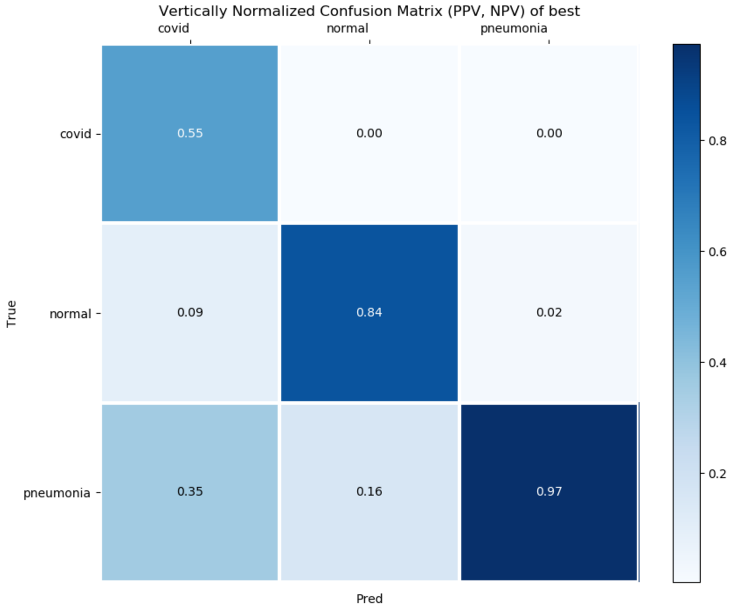

Figure 10: vnorm confusion matrix 2

take a closer look, the model not only has done a very good job in sensitivity for three labels, but also did similarly in precision for three labels. The only loss is precision for COVID-19, and the point lost is mostly due to misclassified as other pneumonia. This would make sense to me as other pneumonia and COVID-19 may share some same medical symptoms in lung areas, making them two harder to separate out, than normal cases.

Does the feature make sense? Here's a Grad-CAM example of the best ResNet-18 on iteration 2 compared to best ResNet-18 on iteration 1 we just discussed:

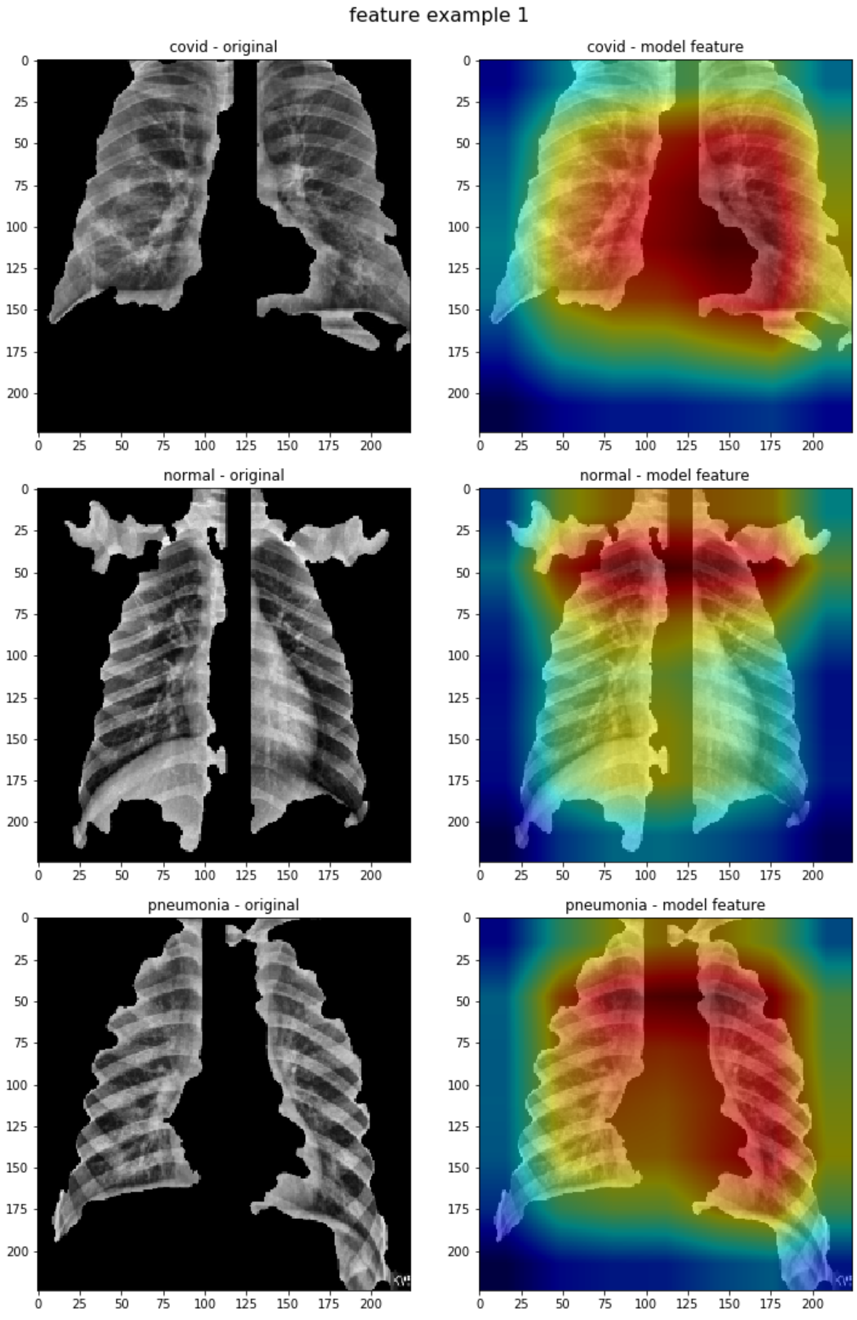

Figure 11: ResNet18.Iteration2

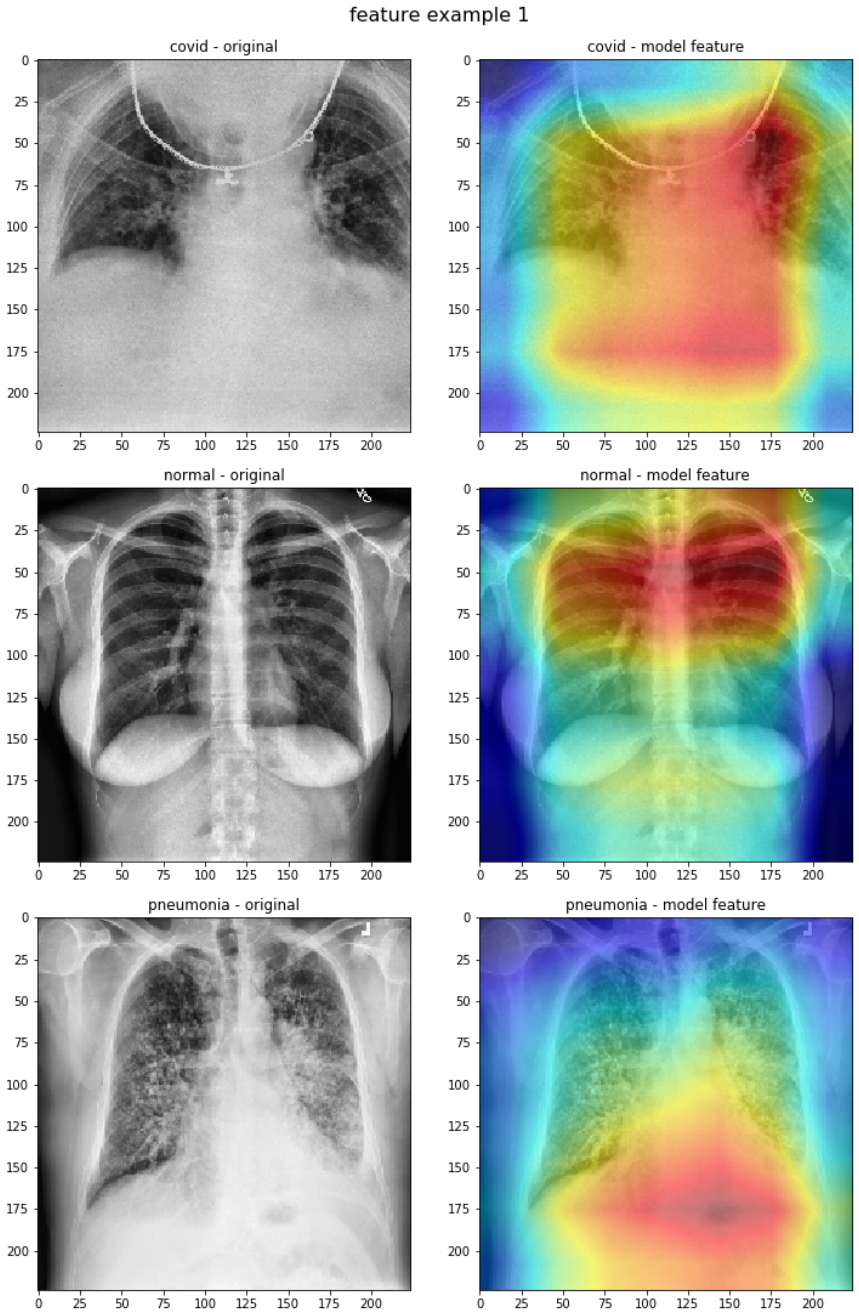

Figure 12: ResNet18.Iteration1

From above two, you can see for the model on iteration 1, features mostly focus outside or on the rim of the lung area, whereas for the model on iteraition 2, features mostly focus in lung areas. Although not stricly that way, this still indicates ResNet-18 on iteration 2's features may be medically valid, at least more valid than the model on teration 1.

Conclusion

We've compared four models' performances through confusion matrix. At this point, it is safe to conclude that, through iteration 1 to iteration 2, we see image resolution, image contrast, key medical area segmentation does play a positive effect on deep neural network ResNet's feature caption and classification performance. As the final best optimal model due to this limited resource dataset, limited time (2 month individual project) and personal computing machine, 0.85

sensitivity, 0.55 precision on target label, with not bad visualization on the feature extraction, considered the different phases of the patients and the validity of Chest X-Ray itself as medical diagnosis for COVID-19, this is not a bad result. Far from a production model, the author does hope, with higher computing power and computing time, better architecture of neural networks, and clearer or higher resolutional topography (maybe CT will be a better disease indicator), the model can someday eventually capture medical patterns with same or better accuracy as radiologists do.

In the mean time, the author proposes this machine learning standardized pipeline graph, with easy one-time setup and one-time run per iteration, for future data science projects of one's own. The source code link is provided at the front of the paper.

Challenges

First challenge is limited computing power due to single local personal machine. One image is trained taken 1 second, 20,000 iamges per epoch is then more than 6 hours. If the network is deeper or has more dynamic parameters, for example, ResNet-150 compared to ResNet-18, then it would take even longer. One way to solve is to consider move the project to could service with GPU computation. With better computer, can try more complex models such as TResNET[9].

Second challenge limited COVID-19 cases. Although the author used 20,000+ dataset, it is quite unbalanced with only 617 COVID-19 samples. With more data, the model will almost likely be trained better.

Last is since the Linda et al team only provides tensorflow metagraph and checkpoints for their COVID-Net model instead of the whole model structure, there's not way to visualize its features.

References

1. https://github.com/hzhaoc
2. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition, arXiv: arXiv:1512.03385
3. LindaWang, AlexanderWong, 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
4. Maria de la Iglesia Vaya, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltran, Francisco Garca-Garca, Marisa Caparros, German Gonzalez, Jose Mara Salinas, BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients, arXiv:2006.01174v3, 2020-06
5. Gianluca Maguolo, Loris Nanni, A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images, arXiv:2004.12823, 2020-09
6. 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
7. 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
8. https://github.gatech.edu/hzhao341/COVID-CXR#data-source
9. 8. Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, Itamar Friedman, TResNet: High Performance GPU-Dedicated Architecture, arXiv:2003.13630
10. Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015