Thoughts on COVID-19 Pneumonia Detection in Chest X-ray Images

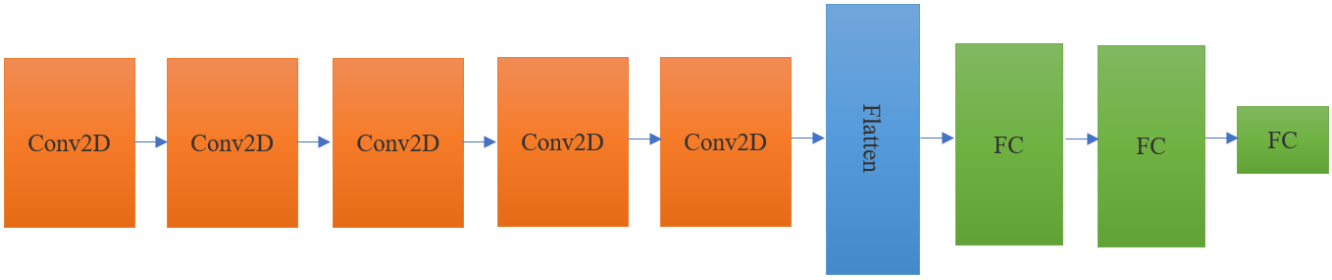
(image from Simon Fraser University News)

When the pandemic was declared in mid-March, many researchers became curious about linking their research fields to the COVID-19 in order to publish some hot-topic stuff … including me! Since I was working on medical image processing, I thought that it could be interesting to do something related to this new trending topic. My first motivation was when I read Adrian Rosebrock's post in PyImageSearch. In that post, he gathered a small dataset of chest x-rays (as of now, we abbreviate it to CXR, which stands for chest x-ray radiographs) and used TensorFlow (Keras) to detect COVID-19 in these CXR images. Following this initial spark, I searched more about articles on the same topic. Ended up finding dozens of posts, articles, and even paper preprints on arXiv and medRxiv claiming high accuracies!

At first, I found the dataset used by the PyImageSearch post: it acquired a dataset of 25 CXRs related to positive COVID-19 cases from a GitHub repo by Dr. Cohen and 25 CXRs of healthy patients from Kaggle's pneumonia CXR dataset. Both were public, and I quickly downloaded them. Kaggle's dataset had normal images as well as bacterial and viral pneumonia ones, related to a classification challenge to detect pneumonia from CXRs. On the other hand, Dr. Cohen's repo was an effort to build an open-source database of images from patients with different diseases; mainly COVID-19, but also there were a few images from SARS, MERS, etc.

Let's Implement a Base Model

I rapidly assembled a couple of convolution layers, followed by some fully connected layers to have my base model. Then, 150 COVID-19 + images and 150 normals ones were collected. As always, I did the training on 80% of data and testing on the rest of them … and viola!



My very simple model with a bunch of Conv2D and FC layers

Model accuracy was about ~97% detecting COVID-19 + from normal. We have got a model with good accuracy and an excellent confusion matrix, so let's publish our work … But wait! Does it really work?! Detecting a chest-related (since now we call it Thoracic because it's a very fancy word!) disease by a model with smaller than 10 layers and less than 500,000 parameters?

In Search of a Reliable Metric

> *The fact was that none of the statistical parameters, like f-score or accuracy, could validate model performance. We needed to actually find out what are the features in the image that the model is deciding based on them? Why an image is classified as COVID-19 +?*

The answer to this critical question is **class activation maps**. We need to probe into a CNN to understand where it is looking to decide on the final class. Simply, we should:

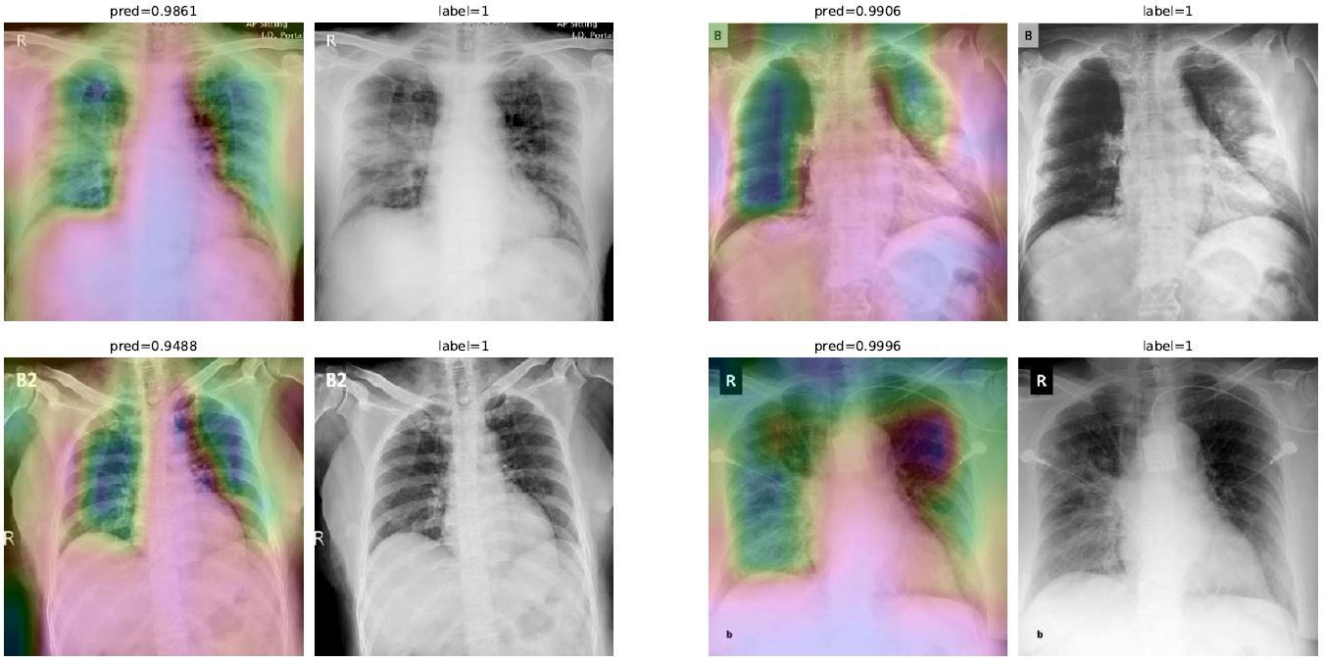
- pass an image through our model
- select the last convolution layer to find out most high-level features of the model
- calculate the gradient flowing into that layer, with respect to the output class for that image

This type of model visualization is called gradient-weighted class activation mapping (**Grad-CAM**). You can search with some keywords to find out more on this: *#class_activation_maps #model_visualization #model_interpritability #visualization_heatmaps*

Let me tell you that I have searched them before and some of the most useful visualization methods are:

- tf-explain, which is defined as a callback function to generate a heatmap every epoch. Perhaps, it will be added to TensorFlow in a future release.
- LIME is another method which divides the input image into components and trains local surrogate models to find regions with the highest impact on the decision.
- CAM-based methods such as Grad-CAM that we will use it here. A very recent method is Score-CAM, presented in CVPR2020. Coding these methods is straightforward, and there are many implementations on the net, such as Keras-vis. One of the bests is Grad-CAM implementation with Keras in PyImageSearch, which we will use it for our task.
- There are other not-very-famous methods, like RISE, that we won't talk about them in this post.

Now that we have our tool, let's see our small basic model's heatmaps for some of our COVID-19 + images:

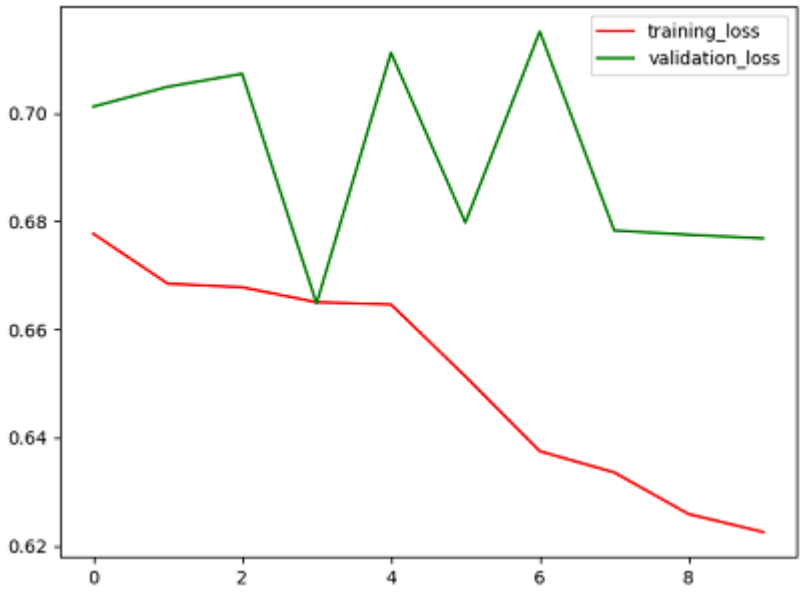


Grad-CAMs of the base model

Thank you my basic model … you are not even looking into the chest lobes … what a desperate result.

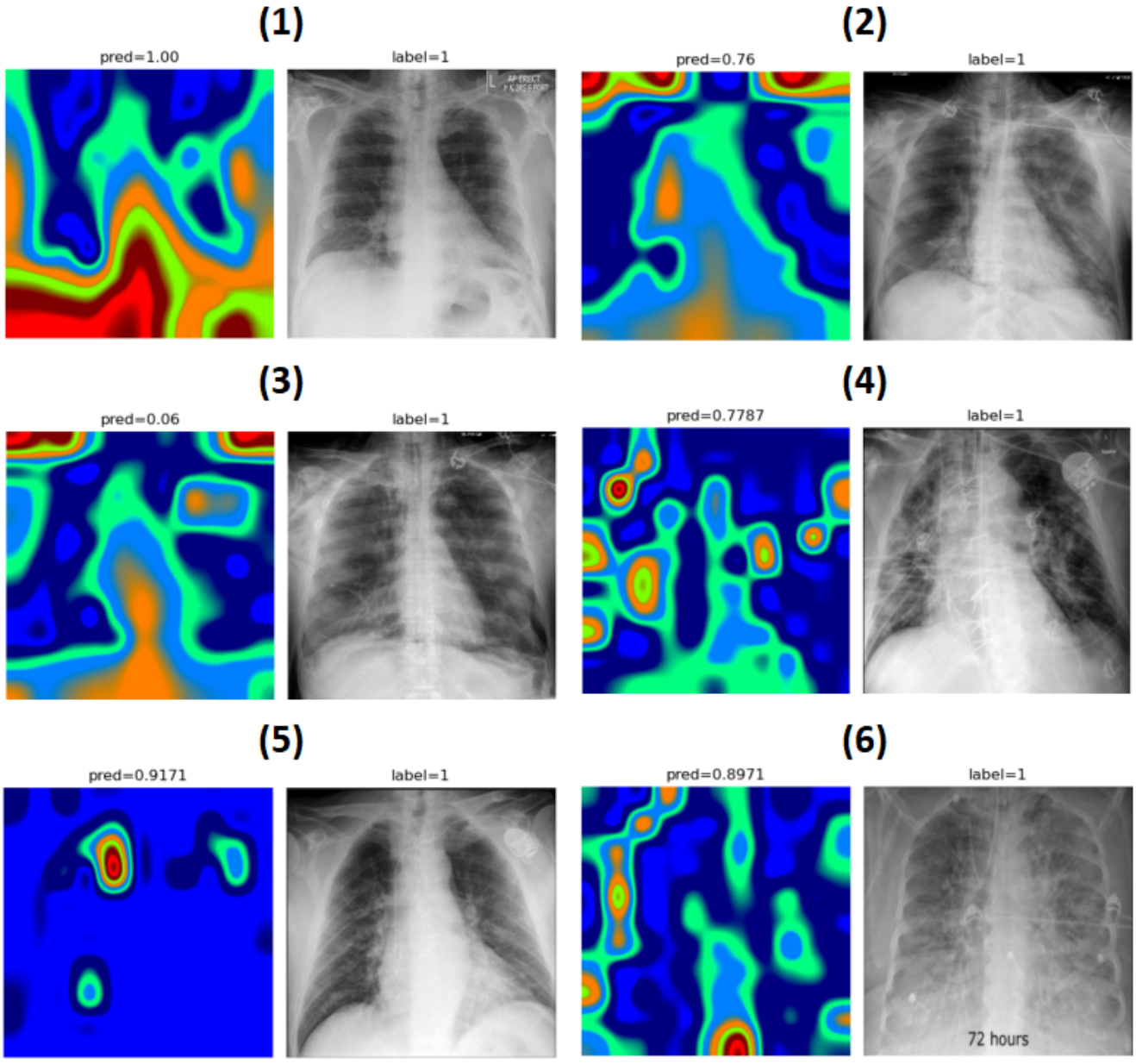
What about Pretrained Models?

Can we publish something like this? Absolutely not. Despite reaching a very high accuracy score, we definitely need a model that categorizes images based on true features. Our base model has the so-called “wrong feature wright decision” problem that needs to be solved. When we have a small dataset, a very fast and efficient approach is **Transfer Learning**. Many articles and papers on COVID-19 detection have benefited from transfer learning, mostly with ResNet, DenseNet and Xception architectures pretrained on the **ImageNet** dataset. Let’s see what happens when we fine-tune a DenseNet on our dataset of 300 images, for example, for 10 epochs. The pretrained model is imported as non-trainable, and then we add a couple of trainable, fully connected layers at the end.



DenseNet-121 fine-tuning curve

We can see that it's definitely not enough. As it consumes a huge resource, we can't let it be trained for many epochs. If trained for hundreds of epochs, it's actually retraining rather than fine-tuning. Why not retraining? We lack enough data for a model with 7M parameters; thus, we definitely face overfitting! So, what happened after a 10-epoch tuning?



Grad-CAMs of DenseNet-121 model, by Keras-viz library

Looking at the images, we can definitely recognize overfitting! Look at (6) to see how the model is looking at the text to classify upon that. On the other hand, in some images such as (5), we can see some regions inside the chest lobes are responsible for the model's decision. Are they correct? We need to know more about COVID-19 pneumonia and its manifestations to better decide on whether our model is working efficiently or not.

COVID-19 Pneumonia Patterns in CXR

It's time for some medical terms. COVID-19 is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). It attacks the respiratory system and results in

pneumonia. Pneumonia itself is also caused by other germs: viruses, bacteria and fungi. Manifestations of lung abnormalities caused by COVID-19 virus are:

- **Ground-Glass Opacities (GGOs)**: they are first signs and very hard to observe in CXRs, but very easy to find in CT scans.
- **Lobar Consolidations**: they mostly appear in lower lobes, and in both lobes (thus, called bilateral).
- **Peripheral airspace opacities**: they are multifocal and easy to detect in CXRs. They also appear as perihilar in some cases as I saw while investigating dataset images and radiologist notes.
- **Diffuse lung opacities**: they are the same in both COVID-19 and other acute respiratory syndromes (I mean SARS).
- **Rare Findings**: They happen in the late stages of disease progression and are uncommon among most cases. We can name pleural effusion and pneumothorax.

Other than the rare findings, others have a huge overlap. All of them are seen as **opacities (whiter than normal)** in CXRs. Let's have a look at an example:

Sample CXR from a COVID-19 positive patient

This CXR is taken from AP erect (lower quality than PA, but higher than AP supine) view captured of a 65 years old male patient admitted with the shortage of breath (SOB, or also called dyspnea) and myalgia (muscular pain). Image findings noted by the radiologist are bilateral ill-defined peripheral airspace opacification in both lungs, normal heart size, and no pleural effusions.

Here are some useful links about pneumonia diagnosis and CXR interpretations:

- Youtube videos for quickly learning how to read CXR images, by Osmosis and MedCram.

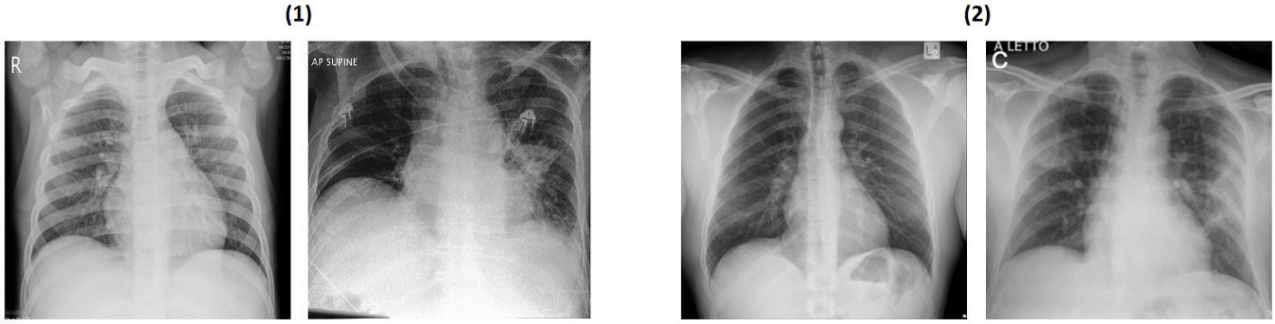
- Radiopaedia article on lobar consolidation detection and COVID-19 detection in CXR/CT.
- An awesome paper about CXR findings of COVID-19.
- A pictorial review on image findings related to COVID-19 with examples for each finding.
- Radiology Assistant article about CXR lung disease detection.

Now that we have learned how to interpret images and diagnose pneumonia (not claiming to become a radiologist! we can just distinguish severe cases from normal ones, not in all images, but most of them), we shall get back to the image (5) in the last figure. Is that a real opacity region detected by the fine-tuned model?

Now that we have learned to have problems even with pretrained models on ImageNet, we may look for probable explanations for this bad performance.

Pediatric Bias

Pediatric means children. The famous Kaggle's pneumonia dataset containing normal cases, as well as several pneumonia types, is captured from children! I asked a radiologist, and he confirmed that characteristics of children's (pediatric pulmonary anatomy) are different from adults in a way that a deep learning model will differentiate them based on the anatomy differences rather than opacities inside the lungs. This is the reason for the red parts of Grad-CAMs, especially those of the base model. Children vs adults in chest x-rays:



Comparison between pediatrics and adults in CXR. Pediatrics in (1) and adults in (2)

Ribs are pretty larger and appeared more horizontally in children. The difference in chest size and the ratio of heart size to the chest width (cardio-thoracic ratio) is noticeable. The moment this bias is confirmed in our dataset, we need to expand it and find more COVID-19 CXRs over the web.

Collecting More CXRs

Earlier I mentioned Dr. Cohen's GitHub repo (ieee8023/covid-chestxray-dataset), which is a valuable open-source project. Hosted CXRs are collected from different sources by scripts. So, I had a look at their sources and also searched for more on the web. Bootstrapped by cloning that repo, I completed my dataset by adding more images from other websites as well. These are the sources:

- **Radiopaedia**: open-edit radiology resource where radiologists submit their daily cases. They have a youtube channel with useful case reports and tutorials on it.
- **SIRM**: the website of the Italian Society of Medical and Interventional Radiology, which has a dedicated database of COVID-19 patients, including both CXR and CT images.
- **EuroRad**: a peer-reviewed image resource of radiological case reports.
- **Figure1**: an image-based social forum that has dedicated a COVID-19 clinical cases section.

- **COVID-19 image data collection**: a GitHub repository by Dr. Cohen et al., which is a combination of some of the mentioned resources and other images.
- **Twitter COVID-19 CXR dataset**: a twitter thread of a cardiothoracic radiologist from Spain who has shared high-quality positive subjects.
- **Peer-reviewed papers**: papers that have shared their clinical images. I have extracted CXRs from nearly 50 papers up until now.
- **Hannover Medical School dataset**: a GitHub repository containing images from the Institute for Diagnostic and Interventional Radiology in Hannover, Germany.
- **Social media**: images collected from Instagram pages. For example, The Radiologist Page or Radiology Case Reports.

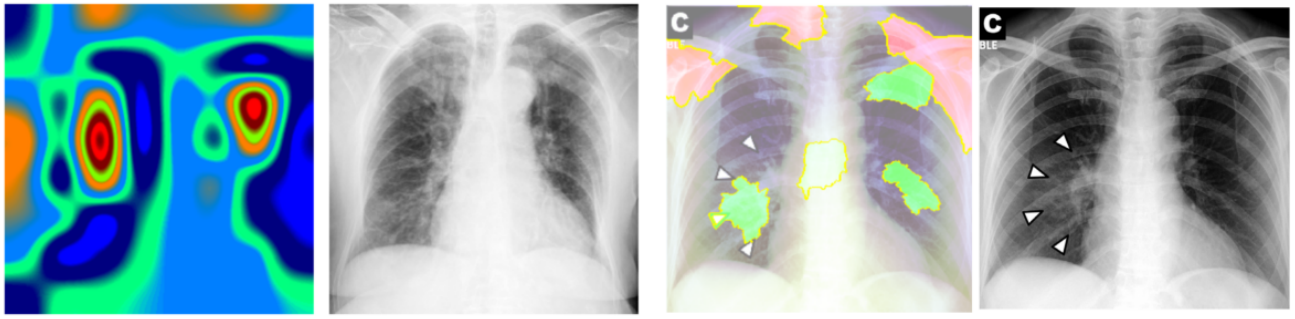
Also, I have found other resources which will talk about at the end! Integrating images from all these sources, we have **~800** images from COVID-19 positive patients. Now we need CXRs from normal chests as well. Normal x-rays seem a bit tricky because you probably won't take x-rays from patients who have no lung diseases. The pediatric image dataset used before has come from Guangzhou Medical Center and includes viral pneumonia, bacterial pneumonia and normal x-rays from one to five years old children. The only large public dataset with (many) normal chest images is **CheXpert** provided by Stanford ML Group.

CheXpert is a large CXR dataset containing more than 224000 images collected from Stanford Hospital. Labels are for 14 different lung diseases, including pneumonia + normal images. As seen below, it has >16000 normal CXRs (labelled as No Finding) and >4500 images labelled as pneumonia (with no further declaration; bacteria, viral or others? we don't know)

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Number of images belonging to each disease in the CheXpert dataset

I found other small datasets having normal CXRs, such as Tuberculosis Chest X-ray Image Datasets that have ~400 normal chest x-rays overall. Now that we have expanded our dataset, I trained my base model on 3400 images (3000 normal and 400 COVID-19). No need to say that accuracy was 98.68% (extremely high again!), here is a model visualization test result:



Heatmaps from the base model trained on 3400 images. Grad-CAM via Keras-viz on the left, LIME visualization map on the right.

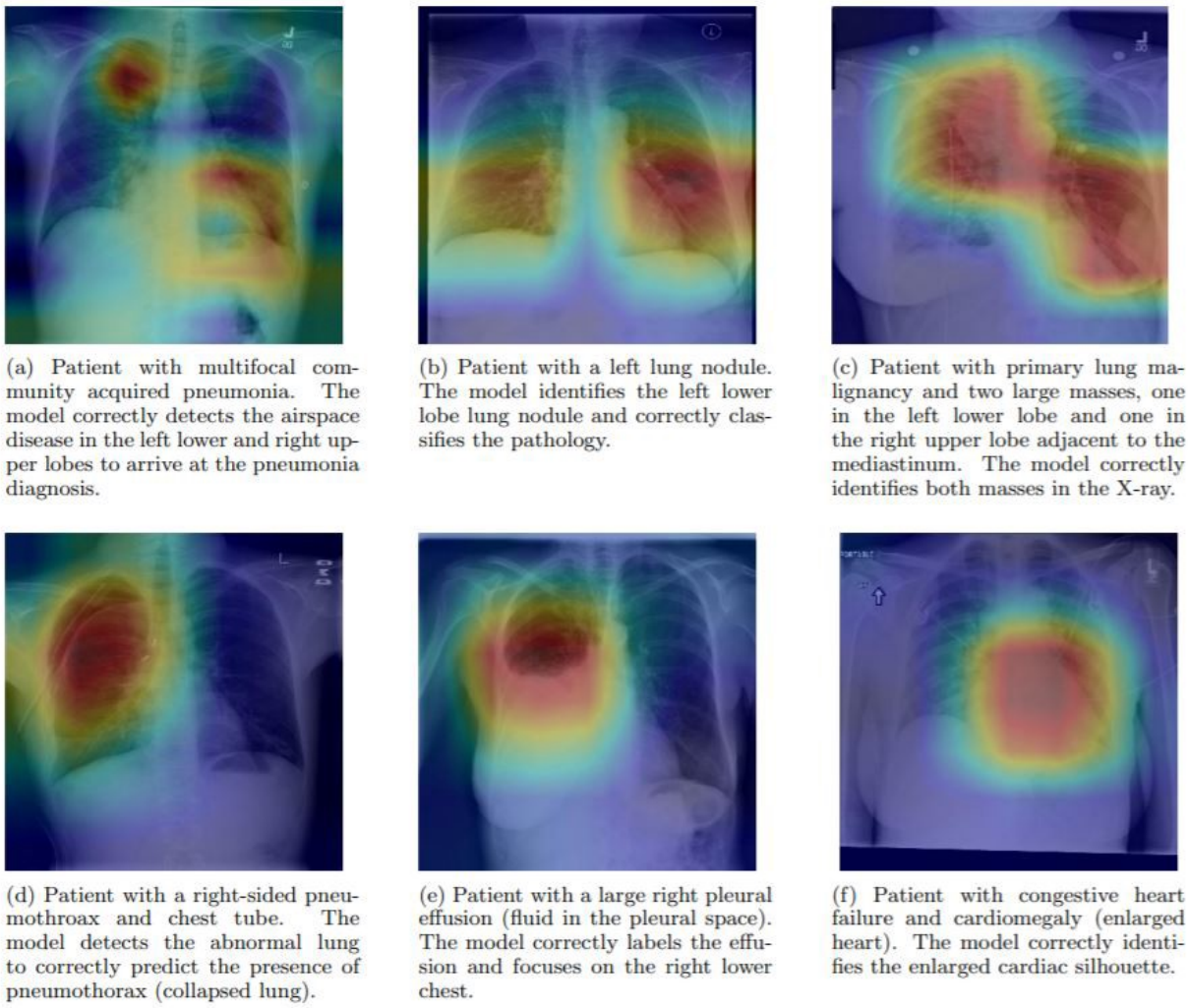
Not bad, but not acceptable! At least it's looking into the lungs. What is that LIME visualization map? It has two superpixel colours: greens that have contributed to the predicted label (positive) and reds that contributed against that label. While one superpixel in the right middle lobe is near the ROI, others have deviated.

Pretrained Models: CheXNet

Up until now, we have figured it out that base models get stuck in a so-called "*right decision with wrong reason*" and ImageNet pretrained models need more data to be retrained for this task. What about a pretrained model on the same type of image data? Remember the CheXpert dataset I mentioned earlier. It was proposed by Stanford ML Group, where they also introduced a model with pretty high metric scores: **CheXNet**.

Authors of CheXNet claimed that they had outperformed radiologists in pneumonia detection on CXRs using deep learning. That seems to be a little bit of hype, in my opinion.

CheXNet is actually a **DenseNet-121** model trained on the NIH CXR-14 dataset of more than 100,000 frontal view x-rays with 14 lung disease labels and also a "*no finding*" label for normal images. Looking through the above tweet by Andrew Ng, you can see that many comments are talking about their certain concerns about the radiologist-level performance claim. I have also found some articles on fundamental problems with CheXNet and its relevant datasets. As an example, this interesting in-depth review has discussed some inconsistencies in different paper versions and some statistical facts about the dataset. On the other hand, visualizations in the paper suggest that it is working quite well in localizing most of the diseases.



(a) Patient with multifocal community acquired pneumonia. The model correctly detects the airspace disease in the left lower and right upper lobes to arrive at the pneumonia diagnosis.

(b) Patient with a left lung nodule. The model identifies the left lower lobe lung nodule and correctly classifies the pathology.

(c) Patient with primary lung malignancy and two large masses, one in the left lower lobe and one in the right upper lobe adjacent to the mediastinum. The model correctly identifies both masses in the X-ray.

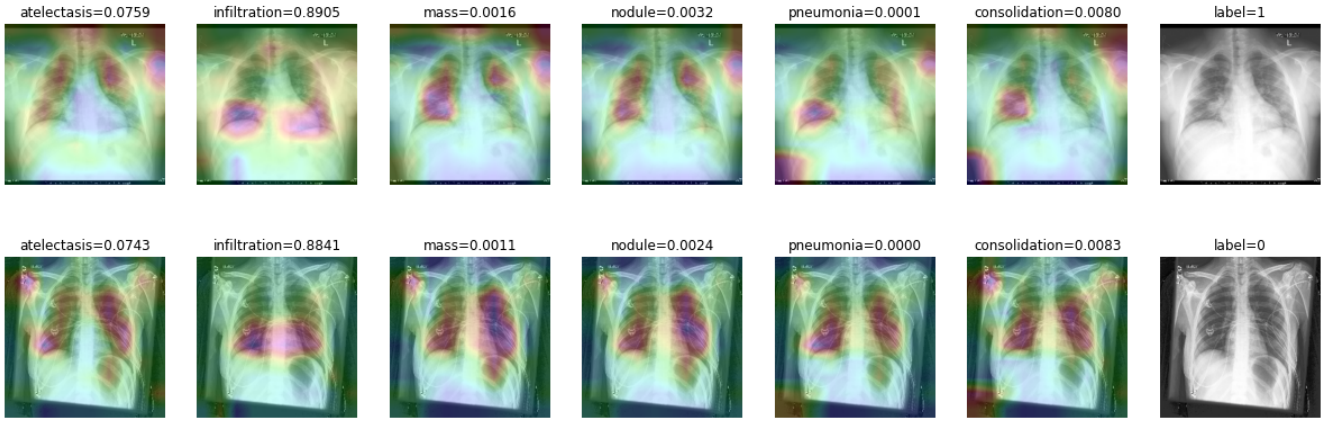
(d) Patient with a right-sided pneumothroax and chest tube. The model detects the abnormal lung to correctly predict the presence of pneumothorax (collapsed lung).

(e) Patient with a large right pleural effusion (fluid in the pleural space). The model correctly labels the effusion and focuses on the right lower chest.

(f) Patient with congestive heart failure and cardiomegaly (enlarged heart). The model correctly identifies the enlarged cardiac silhouette.

visualizations of CheXNet for several labels (directly from the paper on arXiv)

Overall, we were looking for a pretrained network on similar data and now we have it at hand. Now we have two options: whether to use CheXNet as is, I mean just using a combination of model output neurons to apply on our data without training, or to do the fine-tuning. Let's see how it works on some of our images.



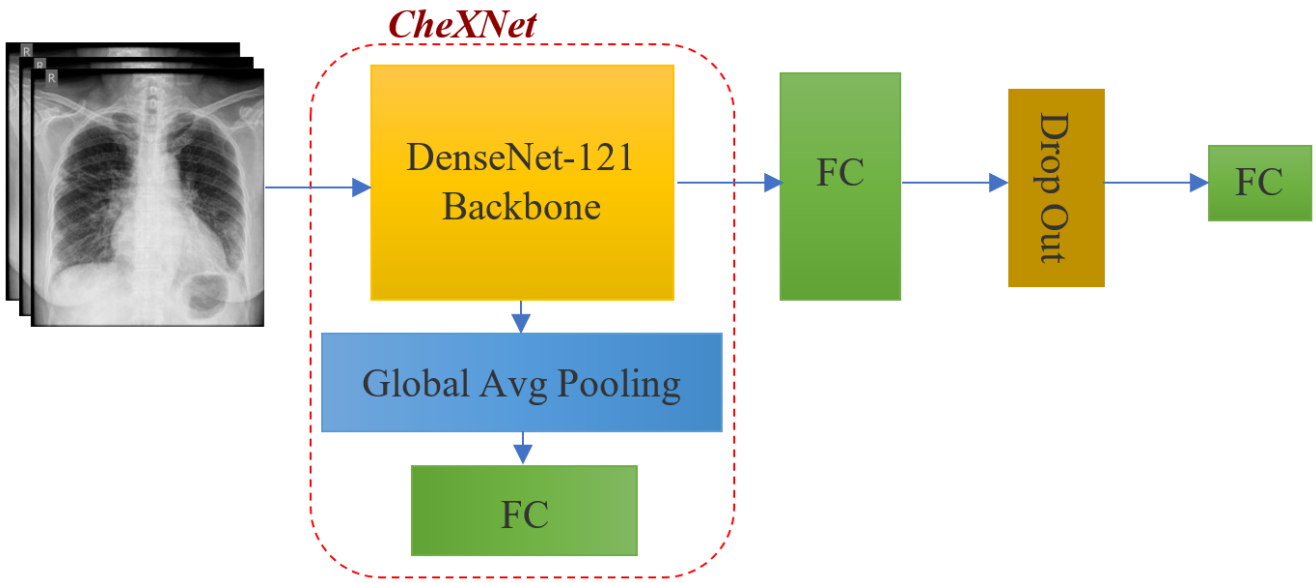
CheXNet probabilities of different classes for a COVID-19 case (top), and a normal case (bottom)

Out of 14+1 output labels, those with the highest correlation with pneumonia are considered to be visualized. The results seem to be acceptable, but there are some problems here. First, infiltration is pretty high for most of the images, while normal CXRs should not have a high infiltration! Second,

there are unwanted absurd regions with high impact in the bottom-left corner in many visualizations. While I cannot guarantee why it happened, I guess it has something to do with a bias in the training database of CheXNet. Thus, we need to go for a fine-tuning and also changing the final dense layers.

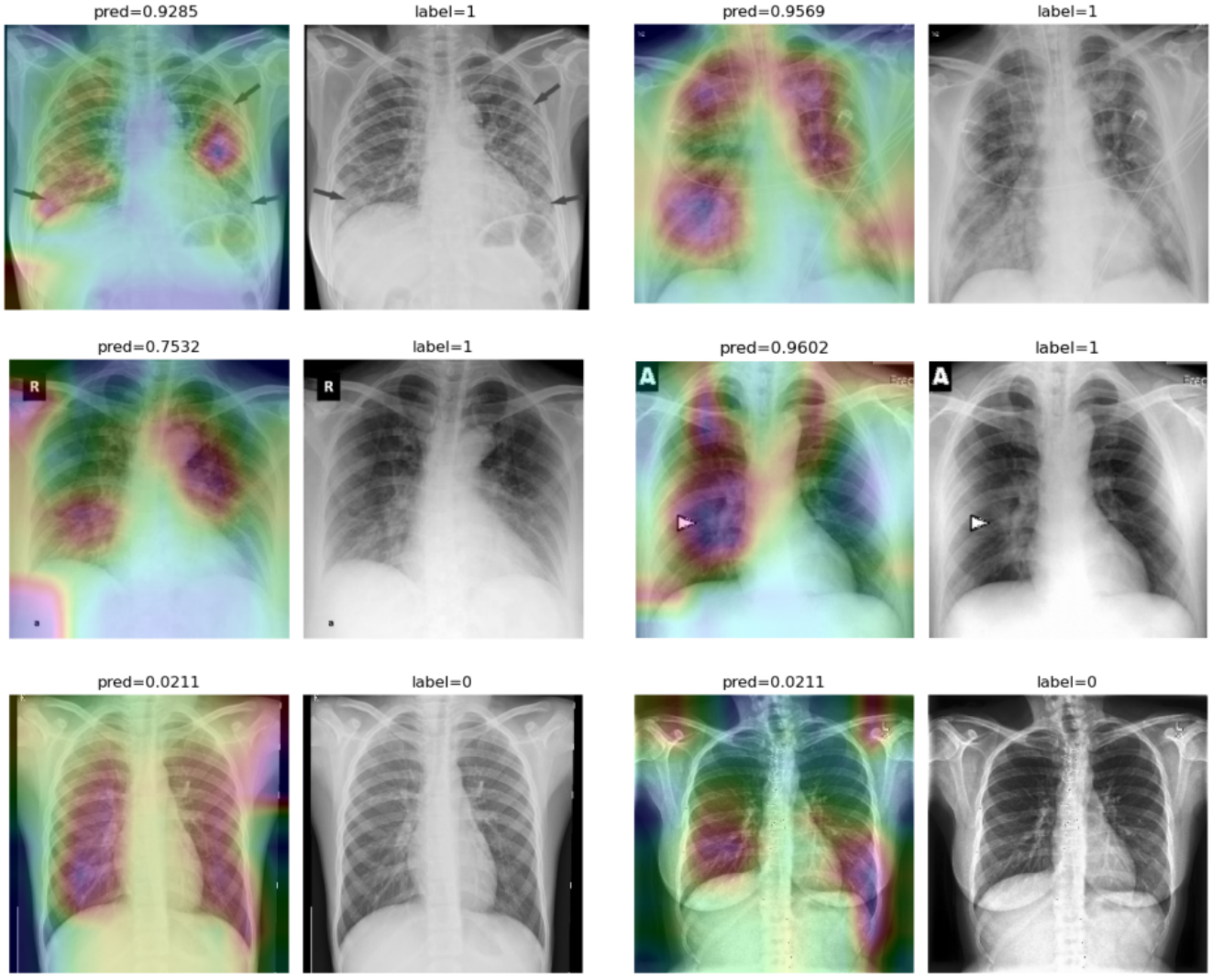
COVID-CXNet

Yeah, I named it as a CheXNet for COVID detection! The architecture is simple: CheXNet's backbone as a trainable block followed by a fully-connected layer with 10 neurons, a drop out layer to prevent overfitting, and a final one-neuron dense layer with the activation function set to Sigmoid.



Our proposed COVID-CXNet architecture

I first tried to use the backbone as non-trainable but didn't get good results after 10–15 epochs. With a trainable backbone and running for 10 epochs, we got a 99.04% accuracy and a 0.96 f-score (for positive class, absolutely!). Dataset was 3628 images (3200 normal + 428 COVID-19) and the test-set ratio was 0.2 of the whole dataset. What about the Grad-CAMs?



Grad-CAMs of our COVID-CXNet

Much better than previous models, eh? The confusion matrix is also like this:

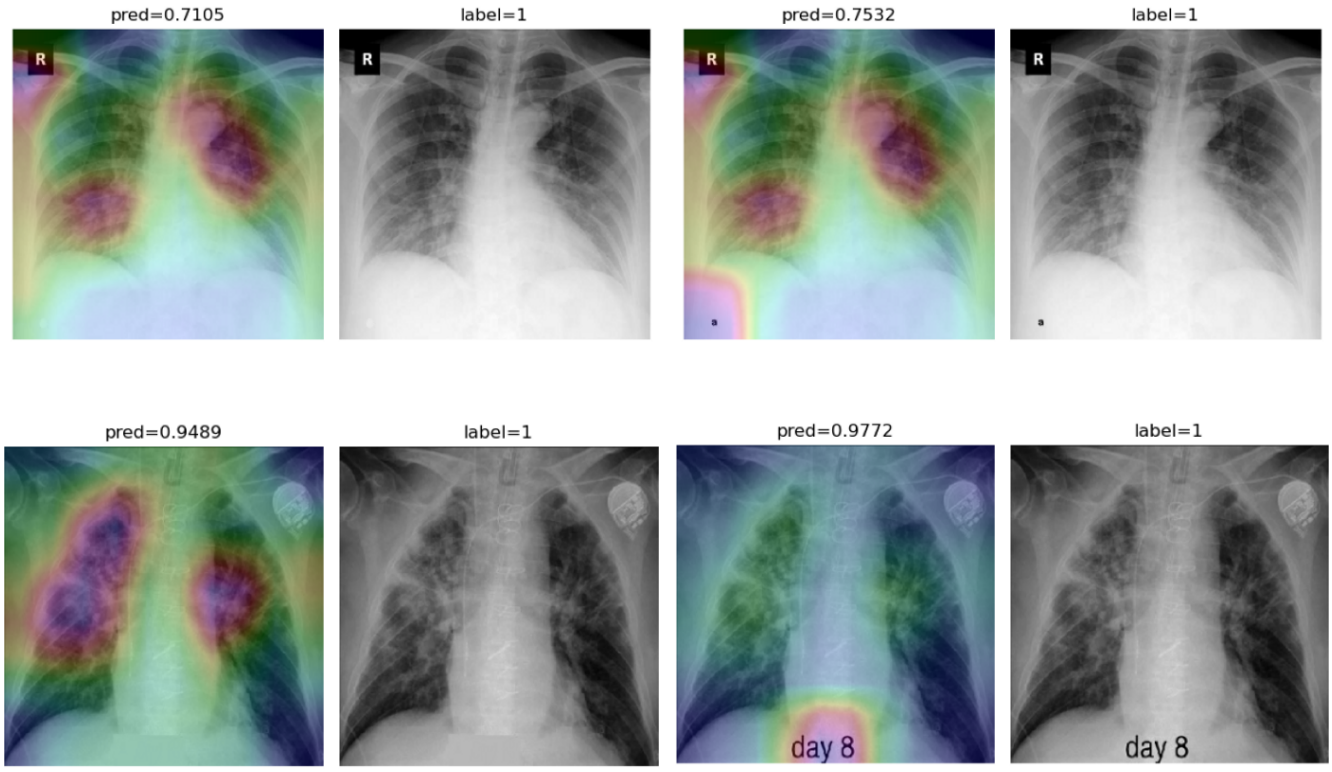
COVID-CXNet		**Predicted**	
		Normal	*COVID-19*
Actual	*Normal*	641	4
	COVID-19	3	78

Confusion matrix of our COVID-CXNet

We reached a pretty high accuracy again, but with better visualization maps. But wait … there are still some problems here. Look at the heatmaps again, can you notice what is wrong? While it is working quite well…

- In some of the images, our model is looking at some irrelevant regions outside the lungs.
- We can see there is still overfitting to frequently-appeared texts and signs, like dates and labels in the corners.

To confirm that the second problem is really related to that sign, I tried removing text and again getting Grad-CAM of the same image:



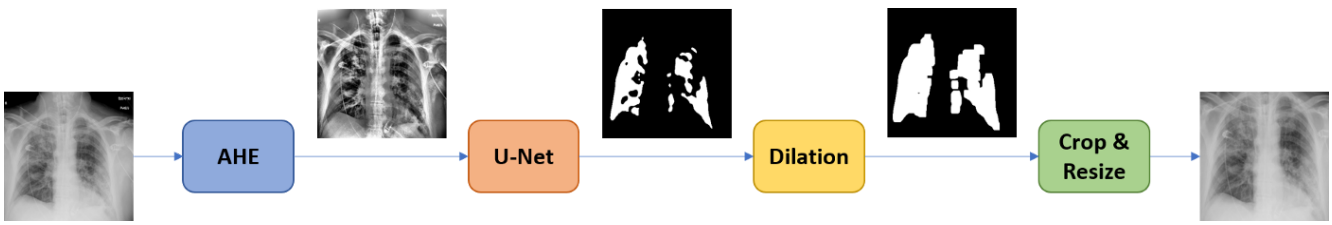
Text removal effect on COVID-CXNet performance; original images on the right, and text-removed images on the left

This definitely has happened because of our small dataset from COVID-19 images. To overcome this problem, the first guess is applying text-removal methods … but can we do something to address both issues mentioned above?

Lung Segmentation is the Key!

By extracting lung regions from the main chest x-ray, we can force our model to look only inside the lungs. To do so, the easiest while the most accurate method is the **U-Net**. U-Net is a fully convolutional network developed mainly for biomedical image segmentation. To utilize a U-Net for lung segmentation, we need pairs of images and their lung-annotated masks. Thus, we need to look for other competitions/research articles carried out on similar CXR data. Hopefully, Shenzhen and Montgomery datasets (previously mentioned as Tuberculosis CXR Datasets in this article) are what we are looking for. There have been different notebooks on Kaggle with details and results. I forked one of them, reran, and used the final hdf5 file for this project.

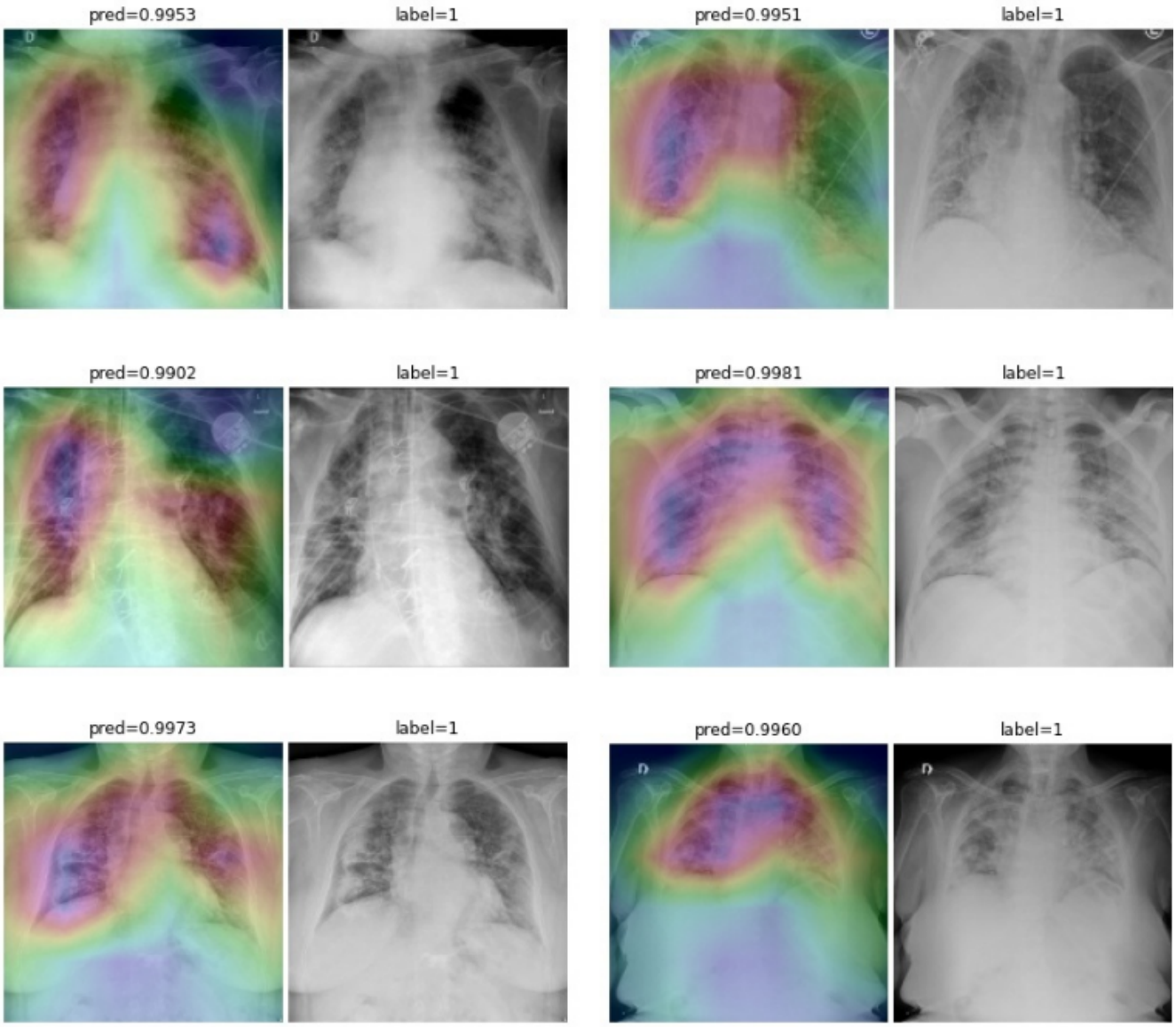
It turned out that exactly grabbing lung regions is not possible because the U-Net segmenter model is trained on different CXRs. So, we can only crop the lung region, trying not to lose any parts of the lung.



U-Net segmentation diagram

This ROI-segmentation block was applied to input images just before applying enhancements, as the first preprocessing stage. COVID-CXNet (with a lung-segmentation block) visualization results can be

seen below:



Grad-CAM visualizations of ROI-segmented COVID-CXNet

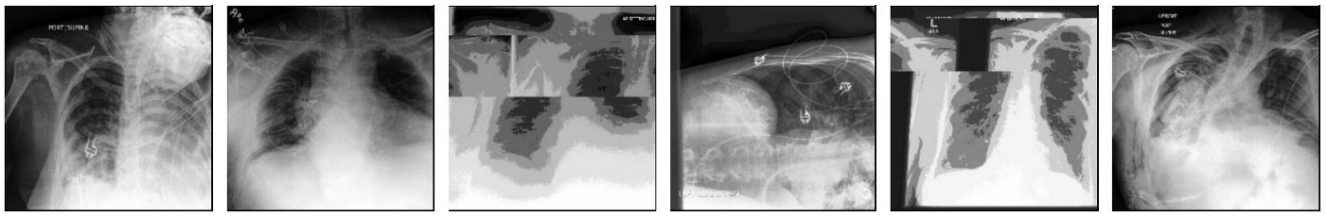
COVID-CXNet is not looking into irrelevant out-of-lung regions anymore. We also had a drop in metric scores in comparison with the previous results: accuracy = 98.62% and f-score = 0.94.

And How about a Multiclass COVID-CXNet?

Most researches have reported three-class classification results; "*COVID-19 Pneumonia (CP)*" vs "*non-COVID Pneumonia (Community-Acquired Pneumonia, abbr. to CAP)*" vs "*Normal*". What is the difference between CP and CAP? CP often has **bilateral** manifestations while CAP is typically **unilateral** (involving only one chest lobe). We must notice that the above-mentioned difference is not always true; in some cases, CP is unilateral, and in several cases, CAP may show itself bilaterally. Now, we need data from CAP images. The only standard non-pediatric dataset, including CAP CXRs, is the CheXpert dataset, to best of my knowledge. I collected ~4800 frontal view CXRs from the pneumonia class in the CheXpert and added them to our dataset with a label of CAP.

In this stage, our system is prone to dataset bias because the majority of images (normal and CAP classes) are mostly from one dataset. Another point about the CheXpert dataset is that a number of

images (I can say ~5%) have various defects. For example, they are: from a lateral view, pediatrics, partly covering the lung region, etc.



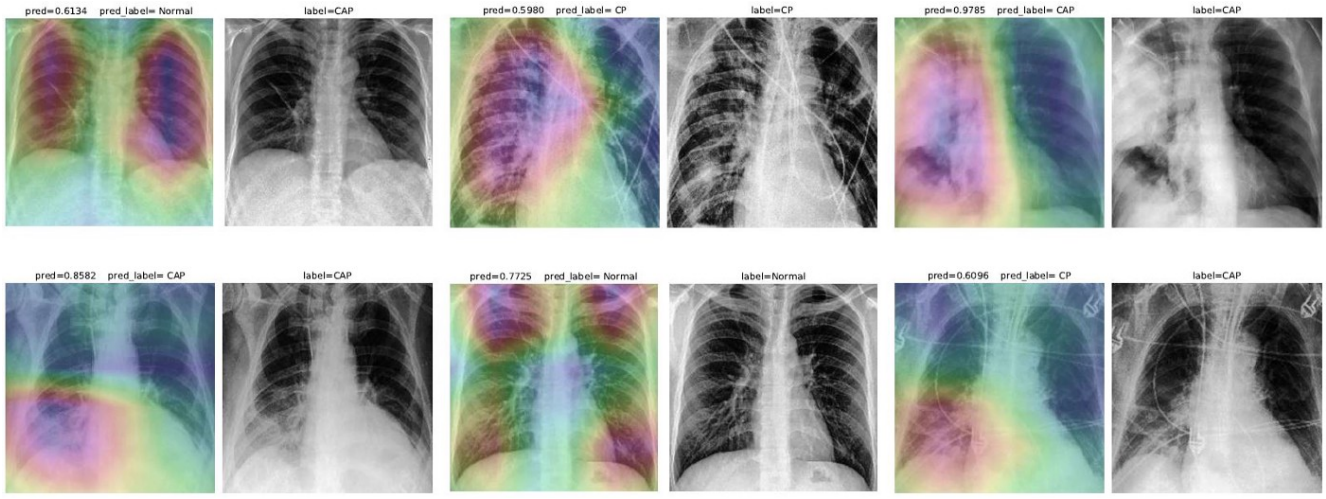
flawed images in the CheXpert dataset

Excluding these images, there are 3500 images left for us. With 3500 normal CXRs and 700 CP images, we have a dataset containing **7700** CXRs to train our multiclass COVID-CXNet. I also benefited from a **hierarchical classification approach** to improve statistical metrics. Firstly, we put CP and CAP together in a "*pneumonia*" class and run a binary COVID-CXNet to classify between normal and pneumonia. Then, we try to differentiate CP from CAP inside that pneumonia class. The confusion matrix is like this:

COVID-CXNet		**Predicted**		
		Normal	*CAP*	*CP*
Actual	*Normal*	689	32	3
	CAP	143	524	5
	CP	3	11	130

Confusion matrix of hierarchical multiclass COVID-CXNet

I personally was waiting to see a huge overlap between CP and CAP, but got surprised to see that overlap between normal and CAP! A probable reason for this phenomenon is that many CAP CXRs are from patients in early-stage pneumonia development. Hence, their lungs might be similar to those of healthy patients. In my opinion, this significant number of false-negatives between CAP and normal classes has something to do with the dataset bias that I pointed earlier. All-in-all, f-score of CP class is 0.92, which seems pretty good. The last step is to see visualizations to confirm our model is doing right.



Grad-CAMs of our multiclass COVID-CXNet

Extracted imaging features show that the model is looking into one chest lobe only to decide on CAP and both lobes to decide on normal or CP labels. There are some flaws like that **overlap between CAP**

and normal in the confusion matrix, or some **wrong predictions** in the sample outputs (above). Since visualizations are acceptable, we can ignore these defects and put the blame on dataset bias! So, these problems are to be solved when we can find more CAP images from different datasets in the future.

> *Now we have built a model to detect COVID pneumonia from non-COVID pneumonia and normal patients. A critical question is that do we really need such a model? Can we use it as a medical decision support system?*

On the Next Level

At this point, a one-word answer is **NO**! A radiologist can perhaps detect COVID-19 pneumonia with higher accuracy. Aside from this fact, is the detection itself the most important thing to be addressed?! COVID-CXNet was only a small step toward building a robust fully automated system. Here, I'm gonna briefly tell you what needs to be done in this field and how we can create a more reliable decision support system.

Larger datasets: As the COVID-19 pandemic has recently happened, there are a few datasets available to the public. While we have collected one of the largest publicly available datasets from different sources I mentioned earlier, many more images are needed for a deep learning model to achieve reasonable robustness. I previously talked about the **CheXpert** dataset with more than 224,000 x-rays with 14 different labels. Another large dataset is the **BIMCV PadChest** dataset constructed of **160,000** CXRs with 19 labels. For COVID-19 x-rays, BIMCV has also introduced **a dataset of 2,265 CXRs**. The (probably) largest open dataset of COVID-19 cases is being developed by RSNA COVID-19 AI Task Force and is named **RICORD**. RICORD is a large dataset of annotated CXR and CT images in DICOM format, and to best of my knowledge is not available yet. But stay tuned as it'll be published very soon!

The role of metadata: A radiologist does not decide based on the CXR itself. He/she will probably look at the patient's clinical symptoms and medical background. Hence, a perfect model needs to have the metadata of each input. Metadata includes various information, such as admission offset, symptoms, medical background, age, sex, etc. We can concatenate a vector of metadata with the input CXR and pass it to the model. Considering metadata to be available, we can go further to find detailed things about a COVID-19 positive patient. For example, we can predict the chance of survival, based on clinical symptoms and the severity of pneumonia features presented by CXR.