COVID19

Early detection in chest images

Informe de “In search for bias within the dataset”

May 12, 2020

Contents

1 Motivation

The experiments performed with chest X-ray images from the Padchest-pneumonia dataset using Deep Learning (DL) models have shown promising results in the classification task that tries to discriminate between patients with pneumonia and controls. However, the moderately low error rates must be examined closely to exclude possible artifacts present in the data or in the experimental design.

With that purpose we have applied some techniques, such as heatmap visualisation, to shed some light on the network's black box. Based on the information provided by these heatmaps, the existence of a significant amount of bias in the dataset is suspected, as many highlighted spots (i.e. relevant regions) of the images did not belong to the lung area. Instead, these heatmaps showed that certain unrelated information was critical for the classification

On the grounds of this previous analysis, we considered that further experimentation was needed to assess the existence of that bias within the dataset (i.e. features not related to the target disease, but correlated with it) as it could be influencing, along with the significant a priori probability (around 75% after some quality filters), the good results.

As detailed in the next section, our networks, trained with the images related to the pneumonia vs normal task, achieved 87.9% accuracy over the set of non-filtered images (i.e. no filters regarding quality or duplicates) of the Padchest-pneumonia dataset. The same network structure applied over the remaining set of images, i.e. after applying the filters proposed in the previous ITI work on this dataset, achieved a 85.3% accuracy.

2 Methodology

As a first step in the investigation, we decided to try to automatically segment the lungs in the images. For this task of lung segmentation, a VGG-11-UNet was trained with some images of the Padchest-pneumonia dataset in which the lungs had been manually segmented.

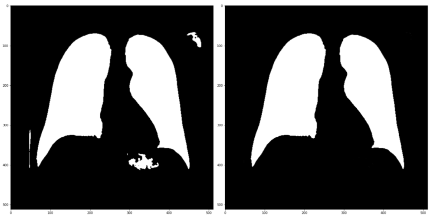

Figure 1: Mask after VGG-UNet (left) and after post-process (right).

Then, masks for lung segmentation of the whole set of frontal RX images were automatically generated by using this network, followed by a post-process (Figure 1). This network achieves 0.9737 DICE and 0.9337 IoU scores over the Montgomery dataset, where DICE and IoU are defined as follows:

$$DICE = \frac{2|A \cap B|}{|A|+|B|} \quad IoU = \frac{A \cap B}{A \cup B}$$

By using this segmentation, some experiments were then performed in order to measure the degree of bias existing in the dataset. The first one was classification over the set of segmented images using the lung segmentation masks dilated 1, 10, 20, 30, 40, 50 and 60 pixels, and eroded 10 pixels, in order to assess how this variability affected the results. When dilating the masks 60 pixels, in most cases the complete RX image is being used.

The second experiment involves training the classifier only from the features outside the lungs. This is performed by occluding the lungs through the mask which segments its area. This mask is dilated 10 pixels to ensure all the information from the lungs is removed. Some image samples can be seen in Figure 2.

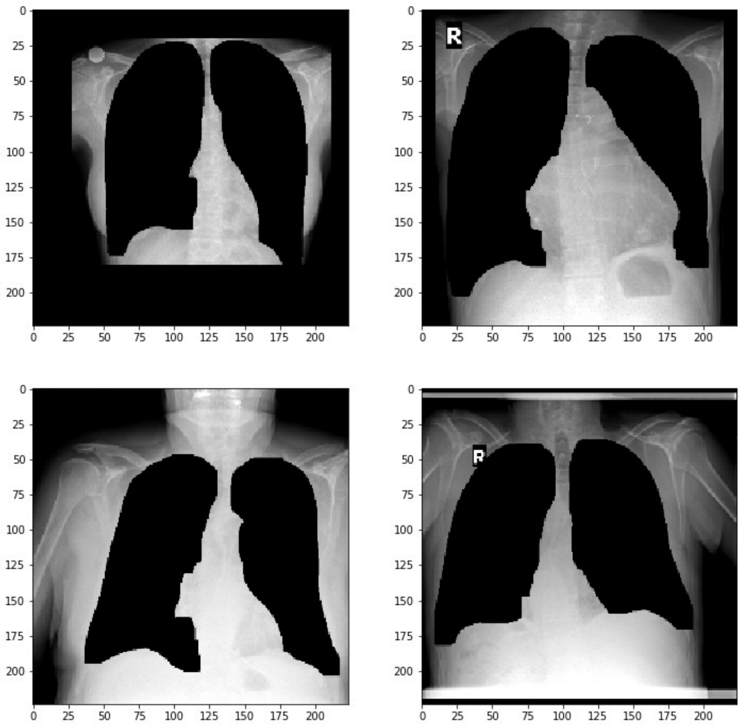

Figure 2: Images without lung info.

As an additional experiment, a classification using only the image metadata, was performed. The subset of the metadata features used was:

- StudyDate_DICOM
- PatientBirth
- PatientSex_DICOM
- Pediatric
- Modality_DICOM
- Manufacturer_DICOM
- PhotometricInterpretation_DICOM
- PixelRepresentation_DICOM

- SpatialResolution_DICOM
- WindowCenter_DICOM
- WindowWidth_DICOM
- Exposure_DICOM
- ExposureInuAs_DICOM
- ExposureTime
- RelativeXRayExposure_DICOM

The experiment was repeated with and without the "StudyDate" feature. For the first and second experiments, a VGG16 network was trained. For the latter, a plain 3-layer perceptron was used.

3 Results

3.1 First experiment

Figure 3 shows how enlarging the segmentation area, the accuracy of the model increases. This suggests a potential bias due to image features outside the lungs. Each reported point is an average of three different experiments with random initialisations. Increasing the number of experiments averaged would smooth the curve. The result of the 60 pixel dilation, which entails using almost the entire image, is equivalent to the one achieved with the whole image, as expected.

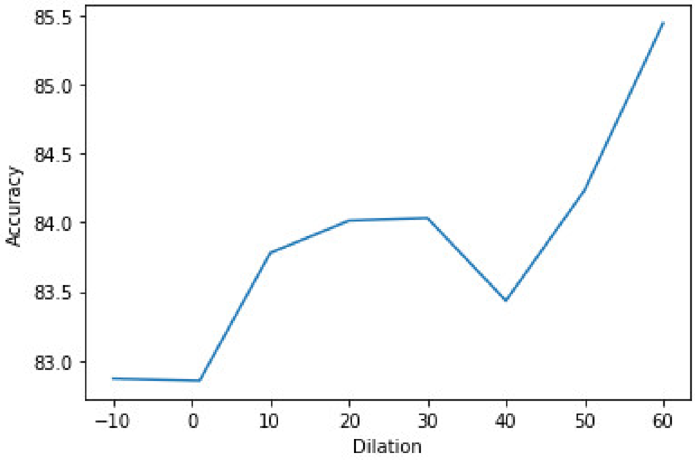

Figure 3: Accuracy as a function of the dilation of the lung masks.

3.2 Second experiment

In this experiment, susprisingly, 84% accuracy was achieved with no pneumonia information within the images whatsoever, as this information should only be found inside the lung region. To further restrict the information supplied to the network, we doubled the dilation applied to the mask and still achieved 83.3% accuracy. As can be seen in Figure 4, the network focuses on surrounding areas which can only contain artifacts and hints that allow the network to learn the bias of the dataset.

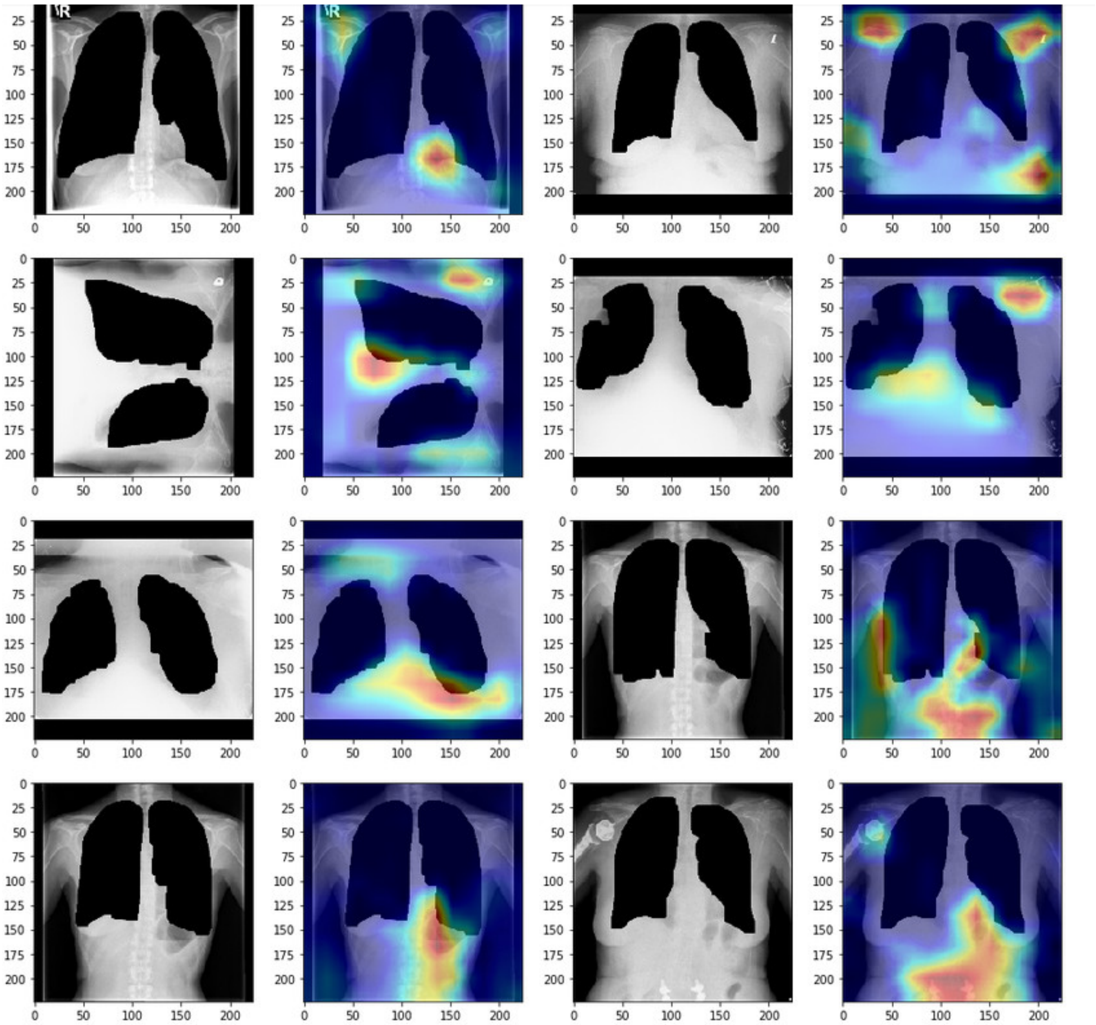

Figure 4: Heatmaps which highlight the areas that influence most the classification.

3.3 Third experiment

Finally, the experiments classifying the metadata achieved 84% accuracy including the “StudyDate” feature and 80% without it. This shows that the date of the capture is critical for the classification of the metadata. It is probable that the time when the RX was acquired causes particular differences in some features of the images correlated to the class, i.e. introducing biases that the networks are exploiting to classify the images. Either way, using only the metadata from the other features, which are definitely related to the quality, levels or textures (or other features) of the images, a 5% increased accuracy over the a priori probability has been achieved.

Particularly, the birth date of the patients has an influence on the lung size in the image. In Figure 5 it can be seen that there are more pneumonia cases than normals in both tales of the distribution. If the patient was born after 2005, the RX could be pediatric and therefore have different geometry, size and features outside the lungs.

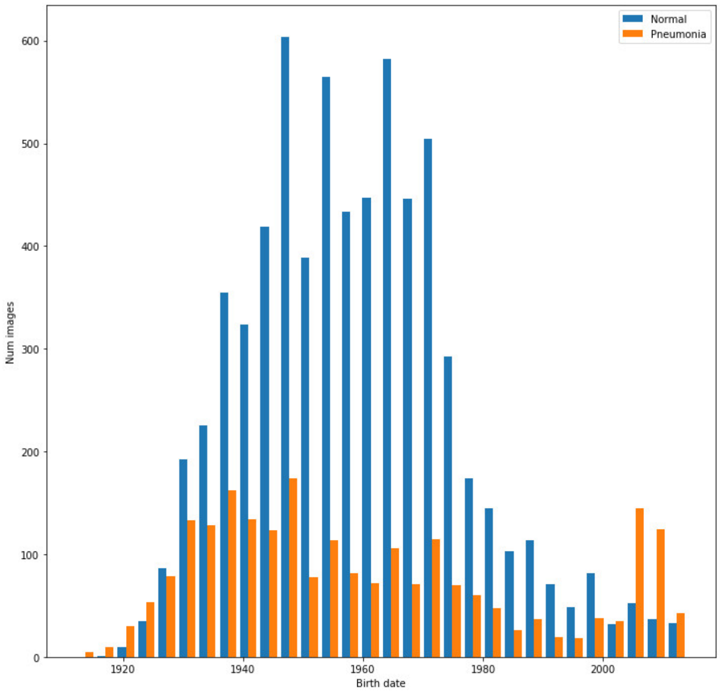

Figure 5: Distribution of the birth date of the patients.

4 Conclusions

Based on the information from the previous experiments we can conclude the existence of a non-negligible amount of bias in the dataset. Therefore, classification results should no longer be reported without removing the noisy information of the surroundings of the lungs. Lung segmentation must be performed in order to achieve meaningful results. Also, features such as the age of the patients, the equipment used, the date, etc. should be balanced in the final dataset to avoid the bias effects that we have detected in the current data. We plan to conduct a study like this with the Covid data when it is available, before reporting any diagnostic accuracy result whatsoever.