Classification and Segmentation of Pneumothorax from Chest X-rays using Deep Convolutional Neural Networks

Arman Haghanifar

October 2020

1 Dataset Details

CXR dataset of this research study is SIIM-ACR Pneumothorax Segmentation dataset which is made available as a Challenge on the Kaggle platform [1]. Image data is in Digital Imaging and Communications in Medicine (DICOM) format, while bounding boxes are stored in a CSV file. DICOM is a format that includes metadata, such as patient sex and age, along with pixel data (image itself) attached to it. A sample CXR metadata, extracted from DICOM file, is shown in Fig. 1.

```
storage info:
study ID: 1.2.276.0.7230010.3.1.2.8323329.1000.1517875165.878026
series ID: 1.2.276.0.7230010.3.1.3.8323329.1000.1517875165.878025
image ID: 1.2.276.0.7230010.3.1.4.8323329.1000.1517875165.878027
storage type: 1.2.840.10008.5.1.4.1.1.7

patient info:
ID: 17d405a3-a0d2-4901-b33a-63906aa48d9f / sex: M / age: 38

x-ray info:
modality: CR / body-part: CHEST / view: PA

image info:
image size: 1024 * 1024
image pixel spacing: [0.168, 0.168]
```

Figure 1: Metadata of a CXR image stored in DICOM format

As seen in the above figure, CXR images are considered as frontal x-rays taken from the chest, with a view position of either Anterior-Posterior (AP) or Posterior-Anterior (PA) projections. Patient names are anonymized and not visible in the dataset. Patient sex and age are also included in the metadata. All images are in a size of 1024×1024 pixels with capturing modality of either Computed Radiography (CR) or Digital Radiography (DR). Next, a random batch of images is selected to visualize how CXRs look like. Fig. 2 shows a random batch of x-rays, where it includes lungs, surrounding bones, and background which is appeared totally black.

[1] https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/

Figure 2: A random batch of 5 images from the dataset

Annotations are in the form of image IDs accompanied by Run-Length Encoding (RLE) masks. RLE is a lossless data compression method in which runs of data are saved as a single data value and count. RLE form is used for images with Pneumothorax mask values where pixel locations are measured from previous ends of the run. Thus, a function must be written to convert RLE-coded masks into mask images. On the other hand, images without Pneumothorax have mask value of -1, which equals to a completely black image as the mask. An example of the RLE form is shown in Fig. 3.

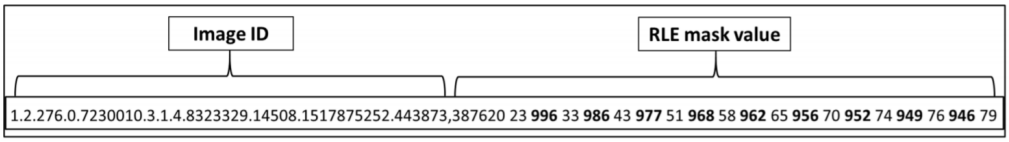

Figure 3: An example of RLE-coded mask

2 Statistical Analysis

The CXR dataset consists of 12,954 binary masks along with 12,047 images in the training-set, and 3205 images in the test-set. Negative images may be from healthy chests or patients with other chest-related diseases, such as Pneumonia. In the training-set, there are more masks than the images. The reason is that some training images have multiple annotations, i.e., multiple Pneumothoraces. In the training-set, there are 9378 images from negative class (77.84%), while there are 2669 ones from positive class (22.16%). Hence, class imbalance is observed in the training-set. Class distribution is illustrated in Fig. 4.

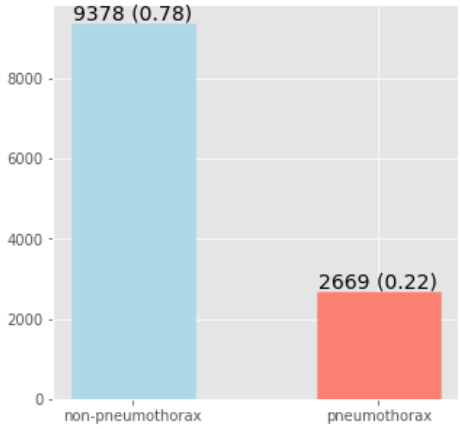

Figure 4: Distribution of dataset classes

In Pneumothorax positive CXRs, 624 have multiple annotation masks and 2669 have only one mask. Among multi-mask images, the most number of masks per image is 10, and most multi-masks have only 2 masks per image. Worth mentioning that multiple masks in an image may overlap each other. Thus, to have one mask as the ground truth for each of the training-set images:

- In negative class images, a black mask is built and considered as the mask for that case.
- In single-masked images, the corresponding mask is the ground truth of the image
- In multi-masked images, all related masks are loaded and the union of them is computed to build the final mask as the ground truth.

Distribution of training-set images in term of patient sex, patient age category, CXR view projection, and CXR modality is shown in Fig. 5 and Fig. 6.

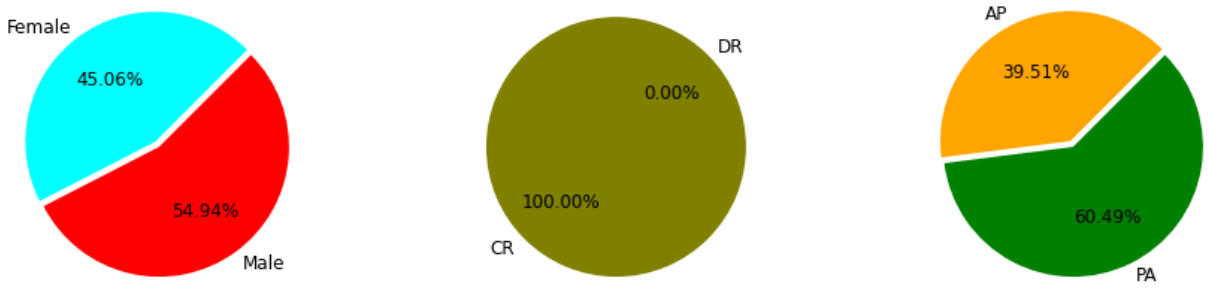

Figure 5: Distribution of dataset images in terms of (1) patient sex, (2) image modality, and (3) image view position

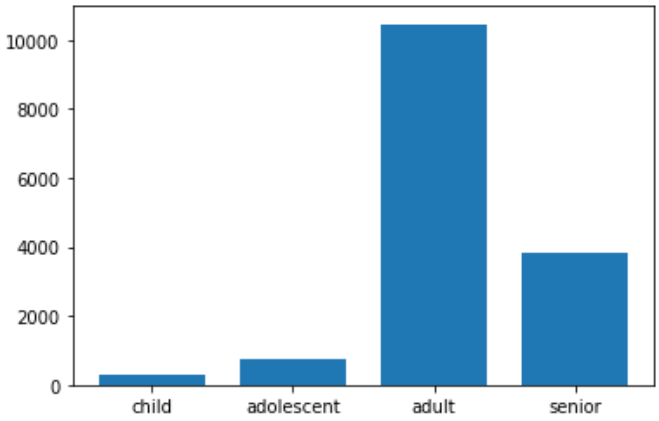

Figure 6: Distribution of images in terms of patient age categories

As seen in above figures, patients are divided into male and female classes with approximately equal ratio. All CXRs are taken with the modality of CR. Majority of CXRs are taken from PA view, which has higher image quality than AP view. More than 80% of patients are classified as adults, while youth patients are quite small. Age category guideline is taken from a CDC project [2].

[2]https://www.cdc.gov/nchs/products/databriefs/db334.htm

3 Problems with the Dataset

An essential part of data science-related projects is to apply various methods to clean the dataset. While there could be many problems with any dataset at first place, Kaggle competition datasets are usually made available after being cleaned. Hence, images in the dataset are from the same size and format. Annotation records all have values of either -1 or sequence of RLE-coded mask numbers.

Common problems with medical datasets are listed below.

- Label noise: Ground truth of medical images are usually prepared by relevant experts; e.g. radiologist in CXRs. There always is a percentage of error with every expert and none has an accuracy of 100%.
- Image type: In this example, there are images from other view positions, i.e. lateral. While images of lateral view of the same patient may be associated with the main AP/PA CXR as a complementary file, they could not be used for model development. By contrast, lateral CXRs can help radiologists detect subtle diseases that are not clearly visible on AP/PA view images, such as pleural effusion.
- Annotations: Boundaries of disease manifestation in images is often not clearly visible. Besides, annotation guideline must be the same for all the images in the dataset. As an example, segmentation of the lung from the chest x-ray could include cardio part (including heart in the lower left lung zone) or not.

After up and running with the dataset and images, an issue is encountered. Dataset description does not clearly indicate the number of images and corresponding masks. When reducing the number of rows in the metadata, by searching for images with multiple masks in the dataset, final number of rows was reduced from 12954 to 12047. Which means that there are 12047 images in the dataset that have at least one corresponding mask in the metadata file. On the other hand, we have 12089 files of DICOM image in the dataset. As a result, there are 42 images with no masks available. These images are investigated by extracting their image IDs, and a couple of them are plotted in Fig. 7.

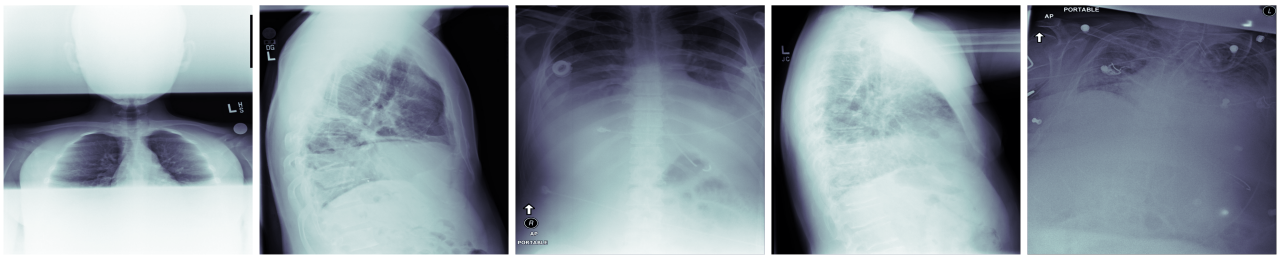

Figure 7: A random batch of 5 images with no corresponding masks in the metadata

As seen above, most mask-less images have various problems; a number of images are from lateral view rather than frontal view, some images do not contain lung area completely or the chest boundaries are not fully seen, etc. Numbers related to the statistical analysis of the dataset mentioned in section 2 is corrected thereupon.

4 Related Public Datasets

There is a lack of sufficient number of x-ray images in the domain of chest diseases. Due to the COVID-19 pandemic during 2020, many research institutes and hospitals tried to make their datasets publicly available. However, small CXR datasets is still a main problem which hinders development of robust chest disease detection models. There is a list of large CXR datasets available online[3], which includes two datasets having Pneumothorax images as part of the data:

- CheXpert[4]: Created and curated by Stanford Machine Learning group, it includes 224,316 chest x-rays from 65,240 patients. There are 14 labels for the images and one of them is Pneumothorax, with 17313 positive and 2663 uncertain cases.
- PadChest[5]: This large dataset is prepared by the Medical Imaging Databank of the Valencia Region (BIMCV) in Spain. Unlike CheXpert, PadChest includes ground truth (report labelling) either created automatically or by radiologists at San Juan Hospital. It includes more than 160,000 images from 67,000 patients. Unfortunately, the number of Pneumothorax cases is not clearly indicated in the dataset statistics.

Besides, there are a number of websites acting as online libraries including cases submitted by radiologists. These cases may also include manual interpretations or even annotation masks prepared by the radiologist. Most cases are images in "jpg" or "png" format and also include metadata such as patient age and sex, reported in the case description. Main online libraries are Radiopaedia[6] and EuroRad [7].

[3]https://github.com/mlmed/cxr-dataset-list
[4]https://stanfordmlgroup.github.io/competitions/chexpert/
[5]https://bimcv.cipf.es/bimcv-projects/padchest/
[6]https://radiopaedia.org/
[7]https://www.eurorad.org/