Capstone 2: Milestone Report

Federico Di Martino

Problem Statement

Medical Image recognition is an established field where data science techniques and especially deep learning with neural networks, are useful. Automating image recognition of diagnostic material can save valuable physician time and improve outcomes for patients.

In this specific case, the problem is to correctly classify whether patients are pneumonic, based on a single x-ray image.

Dataset

The data is Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification from Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification”, Mendeley Data, v2
http://dx.doi.org/10.17632/rscbjbr9sj.2

The dataset comes already split into training, validation and testing sets, with each folder containing NORMAL and pneumonia sub-folders. The total number of images is 5856.

Typical images from the dataset look like this:

Figure 1: An X-ray scan of a pneumonic patient

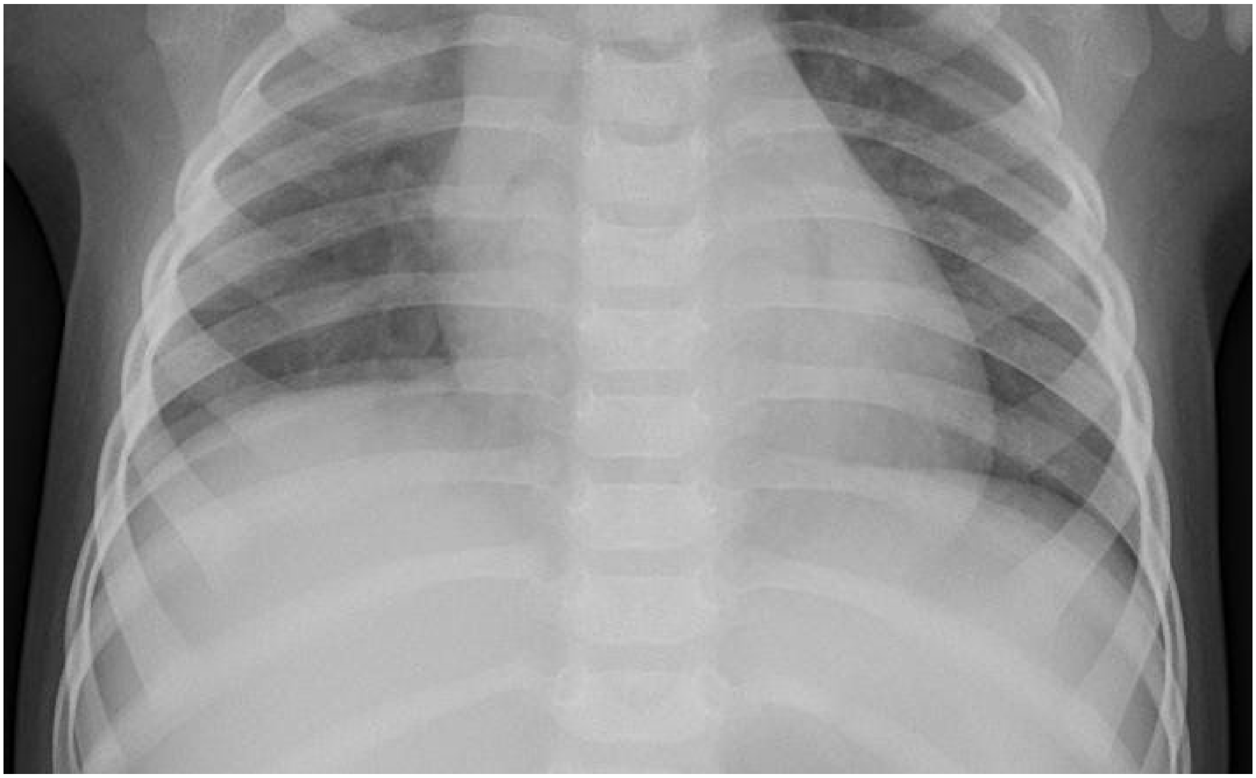
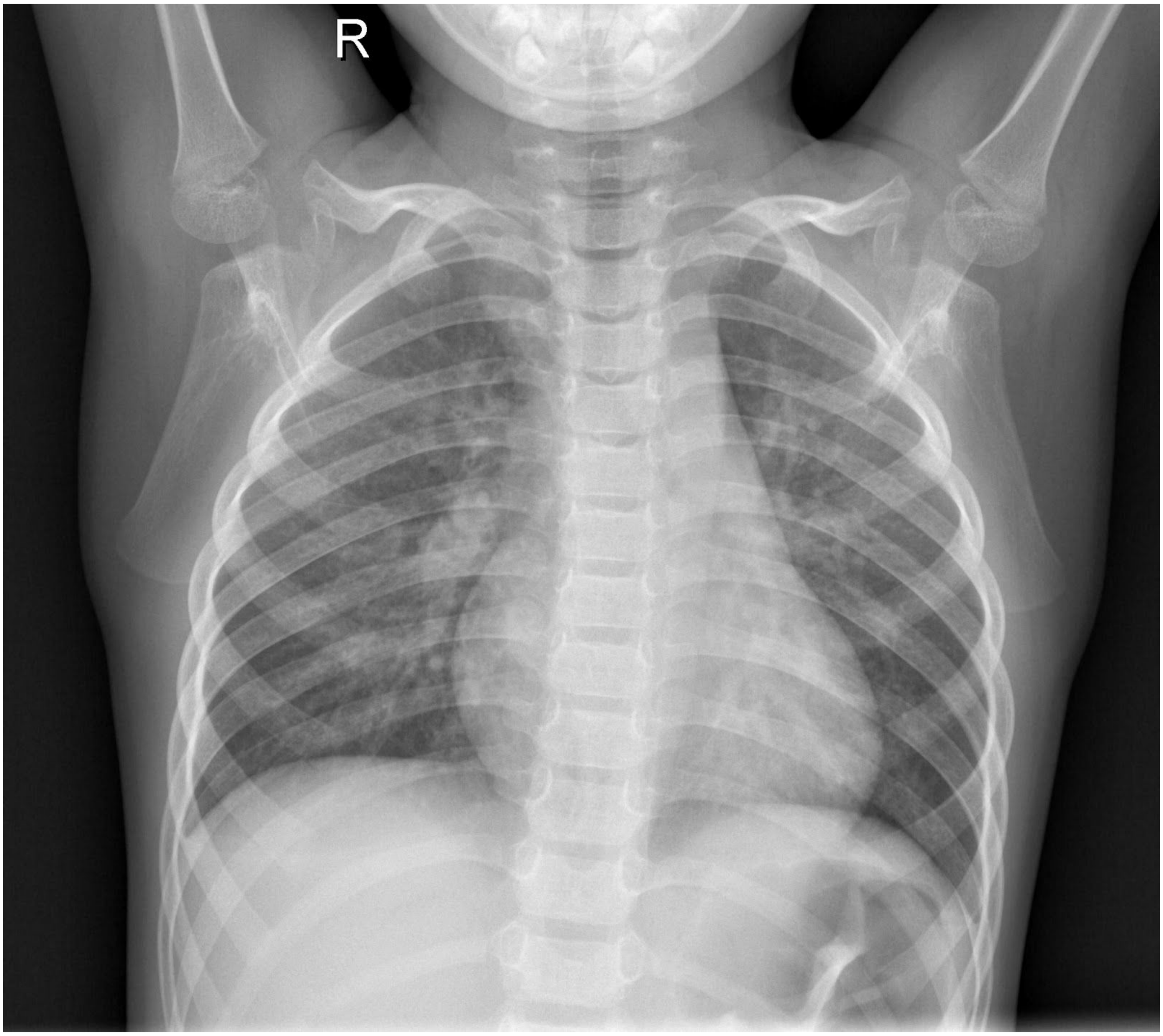


Figure 2: A X-ray scan of a non-pneumonic patient



These are the actual sizes of the images in the dataset and it is important to note that they are different. For a convolutional neural network to work on a set of images they must all be of the same size. I used the PIL library to extract the size information of every image and then resized them all to be the same size as the smallest one. The images are all grayscale but during the resizing process I explicitly converted them all to the same grayscale format (Every pixel represented by a single value for intensity and not RGB).

Here the same images as above resized to the new standard size:

Figure 3: Resized X-ray scan of a pneumonic patient

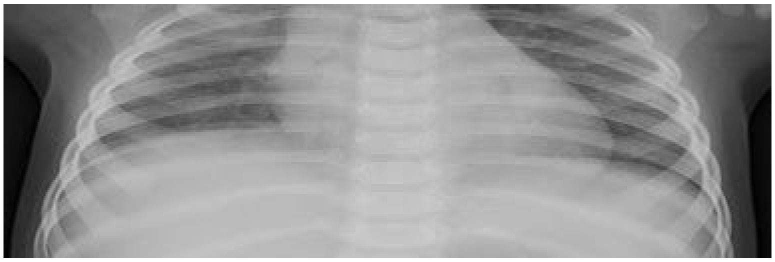
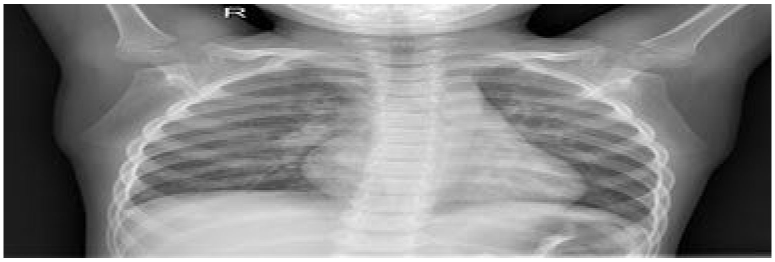


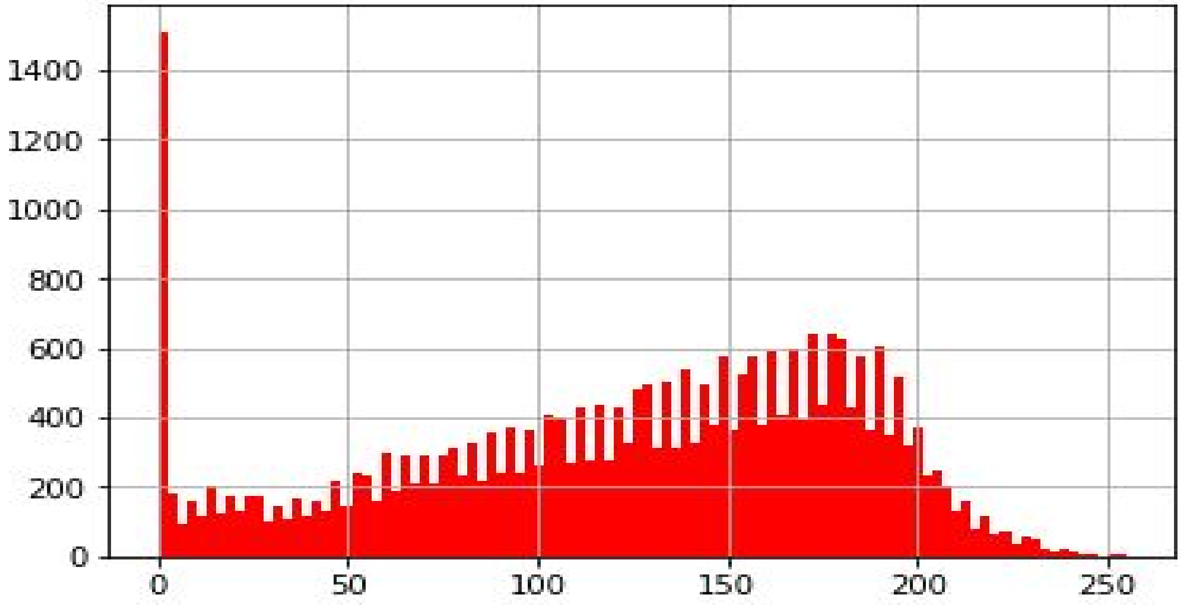
Figure 4: Resized X-ray scan of a non-pneumonic patient.



Initial Findings

The initial analysis I performed was looking at how the pixel intensity is distributed across all the dataset. Using PIL, I extracted the intensity information for every pixel of every image. The result was a list so large that I randomly sampled 0.0001% of it to produce a histogram.

Figure 5: Distribution of pixel intensities in a randomly sampled subset of every pixel.



Apart from the large proportion at 0 (black pixels), there is a relatively even distribution of intensities with few completely white pixels. This fits from what we can see in the example images above.