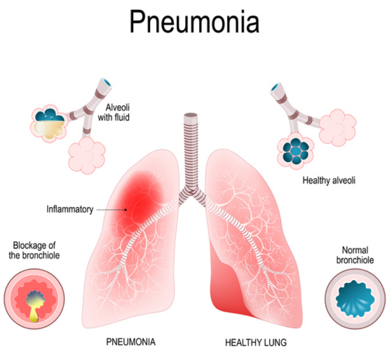
Pneumonia classification

Matej Petrinović

1 Introduction

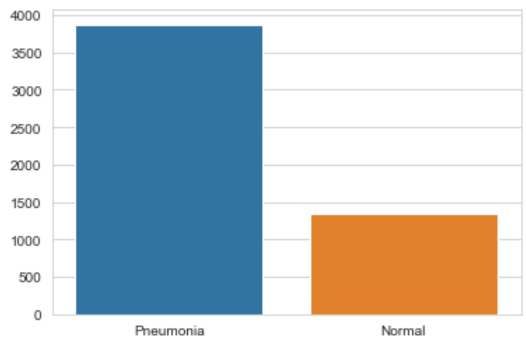
Pneumonia is an infection that inflames the air sacs in one or both lungs. The air sacs may fill with fluid or pus (purulent material), causing cough with phlegm or pus, fever, chills, and difficulty breathing. A variety of organisms, including bacteria, viruses and fungi, can cause pneumonia



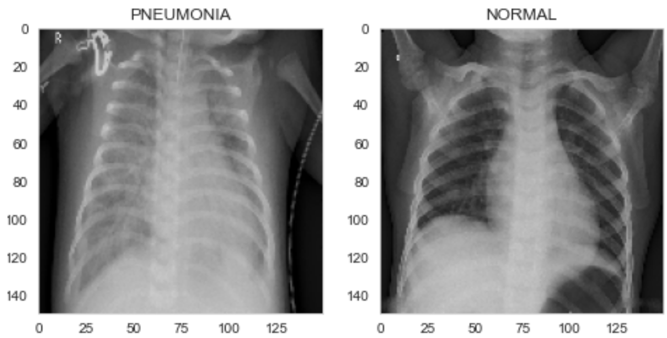
The goal of this work is to make a classification for pneumonia based on X-ray lung scans of the patient. For that a Convolutional neural network will be used.

2 Data Preparation

Before we get started with building a neural network, first we must examine the data. Our data are consists of two classes: Normal Lung Scan and Lung Scan with Pneumonia. Let's take a look on the following graph that shows us the number of pictures in the each class:



As we can see, the data is imbalanced. There are a lot more scans with pneumonia, that scans of lungs without pneumonia. Firsty, take a look of the one scan with and without pneumonia:



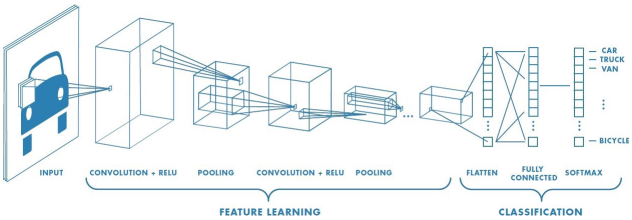
Clearly, we can see that lungs with pneumonia look a some lighter. Because of imbalanced data, we'll do a data augmentation. Data augmentation in data analysis are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. It acts as a regularizer and helps reduce over-fitting when training a machine learning model. It is closely related to oversampling in data analysis. For the data augmentation the following things will be done:

- Randomly rotate some training images by 30 degrees
- Randomly Zoom by 20% some training images
- Randomly shift images horizontally by 10% of the width
- Randomly shift images vertically by 10% of the height
- Randomly flip images horizontally

When the data augmentation is done, we can start with modeling.

3 CNN Model

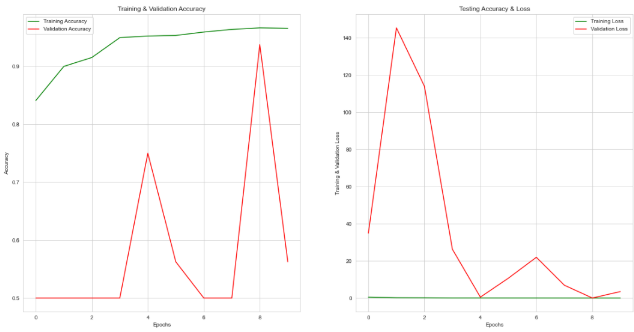
A convolutional neural network is a class of deep neural networks, most commonly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks, based on their shared-weights architecture and translation invariance characteristics.



For activation function we'll use *ReLu* and *Sigmoid* at the and. Loss function we'll be a binary crossentropy function, and an optimizer is RMSprop, a gradient based optimization.

4 Model Evaluation

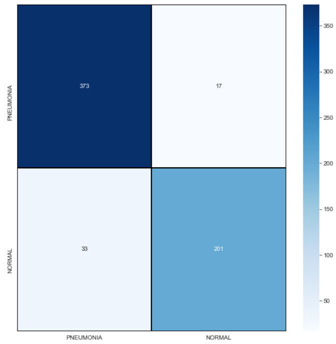
After training the model, the evaluation must be done. Let's take a look on the graph that shows train and validation accuracy and loss over the epochs.



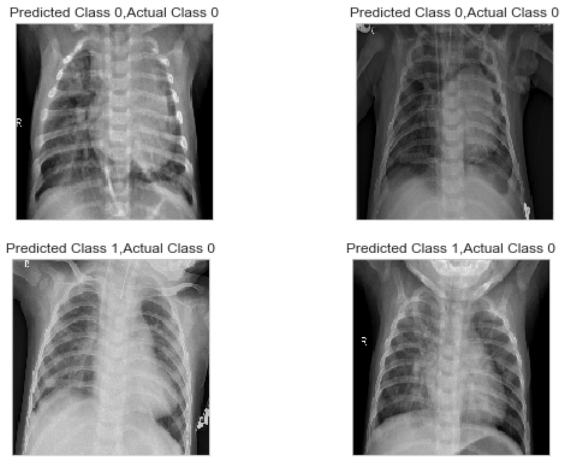
Validation part certainly does not look promising. Take a look on the classification report table for a test data.

	precision	recall	f1-score	support
Pneumonia (Class 0)	0.92	0.96	0.94	390
Normal (Class 1)	0.92	0.86	0.89	234
accuracy			0.92	624
macro avg	0.92	0.91	0.91	624
weighted avg	0.92	0.92	0.92	624

Test data shows some better results, that a validation. This is good, because the model has never seen this data. Confusion matrix is shown below:



From there we can calculate the accuracy, which is $ACC = 0.9199$, and loss is 0.2895. And for the end, let us look some right classification and some wrong.



We can easily see that model predicted pictures without pneumonia as lung with pneumonia, but reason is that as you can see that these lungs look the very similar with the ones with pneumonia.