Capstone Project  
Machine Learning Engineer  
Nanodegree  
Meskhidze Konstantin  
October 6, 2018



Definition

For the final project I’ve decided to join to one of competitions on [Kaggle.com](https://www.kaggle.com/c/home-credit-default-risk).

Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally. In 2015, 920,000 children under the age of 5 died from the disease. In the United States, pneumonia accounts for over 500,000 visits to emergency departments [1] and over 50,000 deaths in 2015 [2], keeping the ailment on the list of top 10 causes of death in the country.

While common, accurately diagnosing pneumonia is a tall order. It requires review of a chest radiograph (CXR) by highly trained specialists and confirmation through clinical history, vital signs and laboratory exams. Pneumonia usually manifests as an area or areas of increased opacity [3] on CXR. However, the diagnosis of pneumonia on CXR is complicated because of a number of other conditions in the lungs such as fluid overload (pulmonary edema), bleeding, volume loss (atelectasis or collapse), lung cancer, or post-radiation or surgical changes. Outside of the lungs, fluid in the pleural space (pleural effusion) also appears as increased opacity on CXR. When available, comparison of CXRs of the patient taken at different time points and correlation with clinical symptoms and history are helpful in making the diagnosis.

CXRs are the most commonly performed diagnostic imaging study. A number of factors such as positioning of the patient and depth of inspiration can alter the appearance of the CXR [4], complicating interpretation further. In addition, clinicians are faced with reading high volumes of images every shift.

To improve the efficiency and reach of diagnostic services, the [Radiological Society of North America (RSNA®)](http://www.rsna.org/) has reached out to Kaggle’s machine learning community and collaborated with the [US National Institutes of Health](https://www.nih.gov/), [The Society of Thoracic Radiology](http://thoracicrad.org/), and [MD.ai](https://www.md.ai/) to develop a rich dataset for this challenge.

In this competition, I’m challenged to build an algorithm to detect a visual signal for pneumonia in medical images. Specifically, your algorithm needs to automatically locate lung opacities on chest radiographs.

Data Description

There are the following files in a dataset:

* **stage\_1\_train\_images.zip** and **stage\_1\_test\_images.zip** – training and test images.
* **stage\_1\_train.csv** **– training labels.**
* **stage\_1\_sample\_submission.csv -** provides the IDs for the test set, as well as a sample of what your submission should look like.
* **stage\_1\_detailed\_class\_info.csv -** contains detailed information about the positive and negative classes in the training set, and may be used to build more nuanced models.

Data fields

* **patientId** \_- A patientId. Each patientId corresponds to a unique image.
* **x**\_ - the upper-left x coordinate of the bounding box.
* **y**\_ - the upper-left y coordinate of the bounding box.
* **width**\_ - the width of the bounding box.
* **height**\_ - the height of the bounding box.
* **Target**\_ - the binary Target, indicating whether this sample has evidence of pneumonia.

All files can be downloaded from <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data>

Solution statement

In the project I will be following the next process:

1. [Data overview](https://www.kaggleusercontent.com/kf/4285484/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..wh3Zv6ehZf2sANfOk-n0NA.ybQsnfTHQ9zTaLvlgTisiVIqiwfam9OZZOhn9iokSNthujq8h8k7LsSRBTMN3ZBZ9BigbU7PlzUlSb0sgaL0ZbzieYAbeJ0Gbo9_L9jEyadvv_EQNBXoLzpHQQxrf4LN9QJHGxp_I570KjwA-z0fRy8Q-wk_VAKpmMi1W6mWRiM.st7uo9-sQ5Qx_7HbTybxkQ/__results__.html#Data_cleaning).

* Example image visualizing.
* Labels and data analysis.
* Drawing boxes.

1. Constructing a simple [CNN](https://www.kaggleusercontent.com/kf/4285484/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..wh3Zv6ehZf2sANfOk-n0NA.ybQsnfTHQ9zTaLvlgTisiVIqiwfam9OZZOhn9iokSNthujq8h8k7LsSRBTMN3ZBZ9BigbU7PlzUlSb0sgaL0ZbzieYAbeJ0Gbo9_L9jEyadvv_EQNBXoLzpHQQxrf4LN9QJHGxp_I570KjwA-z0fRy8Q-wk_VAKpmMi1W6mWRiM.st7uo9-sQ5Qx_7HbTybxkQ/__results__.html#Encoding_data) with segmentation using Keras.

* Resample images.
* Choosing network architecture.
* Train and test the CNN.

1. [Using transfer learning to get better accuracy](https://www.kaggleusercontent.com/kf/4285484/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..wh3Zv6ehZf2sANfOk-n0NA.ybQsnfTHQ9zTaLvlgTisiVIqiwfam9OZZOhn9iokSNthujq8h8k7LsSRBTMN3ZBZ9BigbU7PlzUlSb0sgaL0ZbzieYAbeJ0Gbo9_L9jEyadvv_EQNBXoLzpHQQxrf4LN9QJHGxp_I570KjwA-z0fRy8Q-wk_VAKpmMi1W6mWRiM.st7uo9-sQ5Qx_7HbTybxkQ/__results__.html#Success_method_plot).

* Trying different CNN architectures like : VGG16, RESNET52, Xception, Mask R-CNN, AlexNet ect.

1. Comparing the results and load the best on Kaggle.
2. Making a [conclusion](https://www.kaggleusercontent.com/kf/4285484/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..wh3Zv6ehZf2sANfOk-n0NA.ybQsnfTHQ9zTaLvlgTisiVIqiwfam9OZZOhn9iokSNthujq8h8k7LsSRBTMN3ZBZ9BigbU7PlzUlSb0sgaL0ZbzieYAbeJ0Gbo9_L9jEyadvv_EQNBXoLzpHQQxrf4LN9QJHGxp_I570KjwA-z0fRy8Q-wk_VAKpmMi1W6mWRiM.st7uo9-sQ5Qx_7HbTybxkQ/__results__.html#Conclusions).