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).

Kaggle.com launched a competition RSNA Pneumonia Detection Challenge. In this competition kagglers need to build a model to detect a visual signal pneumonia disease in a medical images of a patient’s chest and locate it. [Radiological Society of North America (RSNA®)](http://www.rsna.org/) and National Institutes of Health Clinical Center provide this competition with datasets. The society hopes this competition would help the diagnosis of pneumonia because the disease kills 15% of patients under 5 years old internationally.

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

The accuracy will be evaluated using a ROC curve.

<https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>

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 in images with the disease.

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 (reduce image dimension)
* Choosing network architecture.
* Train and test the CNN.
* ROC curve accuracy

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).